



EDUCATIONAL and PSYCHOLOGICAL MEASUREMENT

QUARTERLY JOURNAL DEVOTED TO THE DEVELOPMENT AND
APPLICATION OF MEASURES OF INDIVIDUAL DIFFERENCES

Robbins-Monro Procedures for Tailored Testing.
FREDERIC M. LORD 3

Relaxed Rank Order Typal Analysis. LOUIS L. Mc-
QUITTY 33

The Stability Coefficient. EDWARD E. CURETON 45

*Integration of Concepts of Reliability and Standard
Error of Measurement.* JOHN L. HORN 57

*A Measure of Agreement among Subjective Judg-
ments.* K. H. LU 75

*The Interpretation of Regression Coefficients in a
School Effects Model.* ROBERT L. LINN, CHARLES
E. WERTS, AND LEDYARD R. TUCKER 85

*Analyzing School Effects: ANCOVA with a Fallible
Covariate.* CHARLES E. WERTS AND ROBERT L. LINN 95

*A One-Way Analysis of Variance for Single-Subject
Designs.* LESTER C. SHINE II AND SAMUEL M.
BOWER 105

(Continued on inside front cover)

VOLUME THIRTY-ONE, NUMBER ONE, SPRING 1971

Dec 31 1971
Library
8/6/71

Vol. 31
1971
Bureau of Education
(S. C. E. A.)
6.3.81
5641

<i>An Empirical Note on Correlation Coefficients Corrected for Restriction in Range.</i> ROBERT A. FORSYTH	115
<i>A Comparison of Computer-Simulated Conventional and Branching Tests.</i> CARRIE WHERRY WATERS AND A. G. BAYROFF	125
<i>Minimizing Order Effects in the Semantic Differential.</i> ROBERT B. KANE	137
<i>Behavioral Cognition as Related to Interpersonal Perception and Some Personality Traits of College Students.</i> C. M. N. MEHROTRA	145
<i>Vocational Interests and Intelligence in Gifted Adolescents.</i> GEORGE S. WELSH	155
<i>Measures of Ego Identity: A Multitrait Multimethod Validation.</i> FRANK BAKER	165
<i>Dimensions of Psychopathology in Middle Childhood as Evaluated by Three Symptom Checklists.</i> ELSIE E. LESSING AND SUSAN W. ZAGORIN	175
<i>Individual Differences in Diagnostic Judgments of Psychosis and Neurosis From the MMPI.</i> NANCY WIGGINS	199
<i>A Special Review of Buros' Personality Tests and Reviews.</i> FRED DAMARIN	215
ELECTRONIC COMPUTER PROGRAMS AND ACCOUNTING MACHINE PROCEDURES	243
BOOK REVIEW SECTION	297

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is published quarterly, one volume per calendar year, at 2901 Byrdhill Road, Richmond, Virginia 23205. Second class postage paid at Richmond, Virginia and other cities.

Manuscripts should be sent in duplicate to Dr. W. Scott Gehman, Box 6907, College Station, Durham, North Carolina 27708. Authors are requested to put tables and footnotes on pages separate from the text and to follow the general directions given in the *Publication Manual of the American Psychological Association*. Journal titles should not be abbreviated.

Subscription rate, \$14.00 a year, domestic and foreign. Single copies, \$3.50. Back volumes: Volumes XX to the present \$14.00 each; Volumes V through XIX, \$10.00 each. Volumes I through IV are available in a small-print edition at \$3.00 per volume (paper bound).

Orders should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina 27708.



EDUCATIONAL and PSYCHOLOGICAL MEASUREMENT

A QUARTERLY JOURNAL DEVOTED TO THE DEVELOPMENT AND
APPLICATION OF MEASURES OF INDIVIDUAL DIFFERENCES

AKC
Pocket 2
Issue to the
Librarian
200
15/9/71

- Reliable and Valid Hierarchical Classification.* LOUIS
L. MCQUITTY AND JEWEL M. FRARY 321
- Systematic Scoring of Ranked Distractors for the As-
sessment of Piagetian Reasoning Levels.* DAVID H.
FELDMAN AND WINSTON MARKWALDER 347
- Notes on Approximate Procrustes Rotation to Pri-
mary Pattern.* ESKO KALIMO 363
- Communality Estimation in Factor Analysis of Small
Matrices.* EDWARD E. CURETON 371
- A Higher-Order Alpha Factor Analysis of Interest,
Personality, and Ability Variables, Including an
Evaluation of the Effect of Scale Interdependency.*
RICHARD J. ROHLF 381
- Typing Ships with Transpose Factor Analysis.* WIL-
SON H. GUERTIN 397
- Considerations When Making Inferences Within the
Analysis of Covariance Model.* CHARLES E. WERTS
AND ROBERT L. LINN 407

(Continued on inside front cover)

VOLUME THIRTY-ONE, NUMBER TWO, SUMMER 1971

<i>How to Write True-False Test Items.</i> ROBERT L. EBEL	417
<i>A Note on Gaylord's "Estimating Test Reliability from the Item-Test Correlations."</i> JOHN BOWERS	427
<i>The Effects of Forewarning and Pretesting on Attitude Change.</i> GLORIA COWAN AND S. S. KOMORITA	431
<i>A Comparative Study of Five Methods of Assessing Self-Esteem, Dominance, and Dogmatism.</i> DAVID L. HAMILTON	441
<i>The Stability of Individual Differences in Strength and Sensitivity of the Nervous System.</i> FRANK H. FARLEY AND HERBERT H. SEVERSON	453
<i>A Revised Procedure for the Analysis of Biographical Information.</i> WILLIAM H. CLARK AND BRUCE L. MARGOLIS	461
VALIDITY STUDIES OF ACADEMIC ACHIEVEMENT SECTION	465
BOOK REVIEWS	561

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is published quarterly, one volume per calendar year, at 2901 Byrdhill Road, Richmond, Virginia 23205. Second class postage paid at Richmond, Virginia and other cities.

Manuscripts should be sent in duplicate to Dr. W. Scott Gehman, Box 6907, College Station, Durham, North Carolina 27708. Authors are requested to put tables and footnotes on pages separate from the text and to follow the general directions given in the *Publication Manual of the American Psychological Association*. Journal titles should not be abbreviated.

Subscription rate, \$14.00 a year, domestic and foreign. Single copies, \$3.50. Back volumes: Volumes XX to the present \$14.00 each; Volumes V through XIX, \$10.00 each. Volumes I through IV are available in a small-print edition at \$3.00 per volume (paper bound).

Orders should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina 27708.



EDUCATIONAL and PSYCHOLOGICAL MEASUREMENT

A QUARTERLY JOURNAL DEVOTED TO THE DEVELOPMENT AND
APPLICATION OF MEASURES OF INDIVIDUAL DIFFERENCES

Handwritten: K-
Docket & the
issue to the
librarian 226/11/71

- A Factor Analytic Interpretation Strategy.* MARGARET
L. HARRIS AND CHESTER W. HARRIS 589
- A Comparative Study of Some Selected Methods of
Pattern Analysis.* LOUIS L. MCQUITTY 607
- A Measure of the Average Intercorrelation.* EDWARD E.
CURETON 627
- Self-Claimed and Tested Knowledge.* RALPH F. BERDIE 629
- A Significance Test for Biserial r .* EDWARD ALF AND
NORMAN ABRAHAMS 637
- Statistical Control of "Impurity" in the Estimation of
Test Reliability.* K. H. LU 641
- Is There an Optimal Number of Alternatives for
Likert Scale Items? Study I: Reliability and Va-
lidity.* MICHAEL S. MATELL AND JACOB JACOBY. 657
- Validation by the Multigroup-Multiscale Matrix:
An Adaptation of Campbell and Fiske's Convergent
and Discriminant Validation Procedure.* JOHN A.
CENTRA 675

(Continued on inside front cover)

VOLUME THIRTY-ONE, NUMBER THREE, AUTUMN 1971

<i>The Robustness of Tilton's Measure of Overlap.</i> RICHARD S. ELSTER AND MARVIN D. DUNNETTE	685
<i>The Relative Efficiency of Regression and Simple Unit Predictor Weights in Applied Differential Psychology.</i> FRANK L. SCHMIDT	699
<i>True Score Theory: A Paradox.</i> J. O. RAMSAY	715
<i>A Test of the Trait-View Theory of Distortion in Measurement of Personality by Questionnaire.</i> SAMUEL E. KRUG AND RAYMOND B. CATTELL	721
<i>Differences between the Miller Analogies Test Scores of People Tested Twice.</i> JEROME E. DOPPELT	735
ELECTRONIC COMPUTER PROGRAMS AND ACCOUNTING MACHINE PROCEDURES	745
BOOK REVIEWS	779

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is published quarterly, one volume per calendar year, at 2901 Byrdhill Road, Richmond, Virginia 23205. Second class postage paid at Richmond, Virginia and other cities.

Manuscripts should be sent in duplicate to Dr. W. Scott Gehman, Box 6907, College Station, Durham, North Carolina 27708. Authors are requested to put tables and footnotes on pages separate from the text and to follow the general directions given in the *Publication Manual of the American Psychological Association*. Journal titles should not be abbreviated.

Subscription rate, \$14.00 a year, domestic and foreign. Single copies, \$3.50. Back volumes: Volumes XX to the present \$14.00 each; Volumes V through XIX, \$10.00 each. Volumes I through IV are available in a small-print edition at \$3.00 per volume (paper bound).

Orders should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina 27708.



EDUCATIONAL and PSYCHOLOGICAL MEASUREMENT

QUARTERLY JOURNAL DEVOTED TO THE DEVELOPMENT AND
APPLICATION OF MEASURES OF INDIVIDUAL DIFFERENCES

A Theoretical Study of the Measurement Effectiveness of Flexilevel Tests. FREDERIC M. LORD 805

A Short Cut Toward a Submatrix Containing Only "Disturbed" Individuals. LOUIS L. MCQUITT 815

Reliability of Multiple-Choice Tests is the Proportion of Variance Which is True Variance. EDWARD E. CURETON 827

The Probability of Misclassification of Students on Multiple Choice Examinations. WALTER H. CARTER, JR. 831

Nonparametric Item Evaluation Index. STEPHEN H. IVENS 843

Bayesian Techniques for Test Selection. W. PAUL JONES AND F. L. NEWMAN 851

Problems with Inferring Treatment Effects from Repeated Measures. CHARLES E. WERTS AND ROBERT L. LINN 857

Are There Two Extremeness Response Sets? LEONARD V. GORDON 867

(Continued on inside front cover)

VOLUME THIRTY-ONE, NUMBER FOUR, WINTER 1971

<i>Prediction of Individual Stability.</i> GEORGE V. C. PARKER	875
<i>A One-Step Nomograph for the Kolmogorov-Smirnov Test.</i> M. REEB	887
<i>Postexperimental Assessment of Awareness in Attitude Condi- tioning.</i> MONTE M. PAGE	891
<i>A Projective Occupational Attitudes Test.</i> LEROY C. OLSEN AND WILLIAM H. VENEMA	907
<i>The Validity of Measures of Eye-Contact.</i> MARVIN E. SHAW, J. THOMAS BOWMAN, AND FRANCES M. HAEMMERLIE	919
VALIDITY STUDIES OF ACADEMIC ACHIEVEMENT SECTION	927
BOOK REVIEWS	1029
INDEX FOR VOLUME 31	1051

This journal is open to: (1) discussions of problems in the field of the measurement of individual differences, (2) reports of research on the development and use of tests and measurements in education, industry, and government, (3) descriptions of testing programs being used for various purposes, and (4) miscellaneous notes pertinent to the measurement field, such as suggestions of new types of items or improved methods of treating test data. Authors are granted permission to have reprints made of their own articles for their own use at their own expense. Manuscripts should be sent in duplicate to Dr. W. Scott Gehman, Box 6907, College Station, Durham, North Carolina 27708. Authors are requested to put tables and footnotes on pages separate from the text and to follow the general directions given in the *Publication Manual of the American Psychological Association (1967 Revision)*. Journal titles should not be abbreviated.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is published quarterly, one volume per calendar year, at 2901 Byrdhill Road, Richmond, Virginia 23205. Second class postage paid at Richmond, Virginia and other cities.

Publication charges to authors are as follows: \$30.00 per page of running text; \$40.00 per page of tables, figures, and formulas.

Subscription rate, \$14.00 a year, domestic and foreign. Single copies, \$3.50. Back volumes: Volumes XX to the present \$14.00 each; Volumes V through XIX, \$10.00 each. Volumes I through IV are available in a small-print edition at \$3.00 per volume (paper bound).

Orders should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina 27708.

EDUCATIONAL and
PSYCHOLOGICAL



MEASUREMENT

Journal of Educational & Psychological Research
(S. C. E. R. J.)
Date *12-6-71*
File *282*
Bureau *lib*

Editor: W. Scott Gehman

Managing Editor: Geraldine R. Thomas

BOARD OF COOPERATING EDITORS

- DOROTHY C. ADKINS, *University of Hawaii*
LEWIS R. AIKEN, JR., *Guilford College*
HAROLD P. BECHTOLDT, *The University of Iowa*
WILLIAM V. CLEMANS, *Science Research Associates, Inc.*
LOUIS D. COHEN, *University of Florida*
JUNIOUS A. DAVIS, *Educational Testing Service*
HAROLD A. EDGERTON, *Performance Research, Inc.*
MAX D. ENGELHART, *Duke University*
GENE V GLASS, *University of Colorado*
E. B. GREENE, *Chrysler Corporation (Retired)*
J. P. GUILFORD, *University of Southern California, Los Angeles*
JOHN A. HORNADAY, *Babson College*
JOHN E. HORROCKS, *The Ohio State University*
CYRIL J. HOYT, *University of Minnesota*
MILTON D. JACOBSON, *University of Virginia*
JOSEPH C. JOHNSON II, *Duke University*
WILLIAM G. KATZENMEYER, *Duke University*
E. F. LINDQUIST, *State University of Iowa*
FREDERIC M. LORD, *Educational Testing Service*
ARDIE LUBIN, *Naval Medical Neuropsychiatric Research Unit, San Diego*
LOUIS L. MCQUITT, *University of Miami, Coral Gables*
WILLIAM B. MICHAEL, *University of Southern California, Los Angeles*
HOWARD G. MILLER, *North Carolina State University at Raleigh*
ELLIS B. PAGE, *The University of Connecticut*
NAMBURY S. RAJU, *Science Research Associates, Inc.*
BEN H. ROMINE, JR., *University of North Carolina at Charlotte*
KENDON SMITH, *The University of North Carolina at Greensboro*
THELMA G. THURSTONE, *University of North Carolina at Chapel Hill*
HERBERT A. TOOPS, *The Ohio State University*
WILLARD G. WARRINGTON, *Michigan State University*
JOHN E. WILLIAMS, *Wake Forest University*
E. G. WILLIAMSON, *University of Minnesota*

VOLUME THIRTY-ONE, NUMBER ONE, SPRING 1971

ROBBINS-MONRO PROCEDURES FOR TAILORED TESTING¹

FREDERIC M. LORD

Educational Testing Service

WHEN computers are used in the schools for instructional purposes, it is a matter of convenience to use them for measurement purposes also (Turnbull, 1968). The problem of securing accurate measurements is different from the problem of giving effective instruction. This paper is concerned entirely with measurement and not at all with instruction.

In *tailored testing*, an attempt is made to tailor the difficulty of the test items administered to the "ability" of the individual being tested. For most purposes, it will be convenient here to think of the problem of testing or "measuring" just one single individual.

A large pool of items must be available at the start of the testing. The statistical characteristics of these items must be known from earlier testings.

If an examinee answers all n items in a test correctly, we are not able to pinpoint his ability level; for example, we can not tell how he compares with some other examinee who also answers all items correctly. A similar conclusion applies if the examinee does not know the answer to any of the test items. Other things being equal, the best measurement is obtained when the examinee knows the answer to roughly half of the items administered. *In tailored testing we try to choose items for administration that are at a difficulty*

¹ This work was supported in part by contract N-00014-69-C-0017, project designation NR 151-284, between the Personnel and Training Research Programs Office, Psychological Sciences Division, Office of Naval Research and Educational Testing Service. Reproduction in whole or in part is permitted for any purpose of the United States Government.

level that matches the examinee's ability, which we infer from his responses to the items already administered. A convenient, if oversimplified, rule for doing this is that when the examinee gives a wrong answer to an item, the next item administered should be an easier one; when he gives a correct answer, the next item administered should be harder. More complicated rules can be investigated (for example, see Wetherill, 1963; Wetherill and Levitt, 1965), but this will not be done here.

Certain questions still remain to be answered before we can actually start testing:

1. What should be the difficulty level of the first item?
2. How much should the difficulty level be changed after any given right or wrong answer?
3. How should the examinee's responses be scored?
4. How should the effectiveness of various possible procedures be compared?

We are not presently able, either theoretically or practically, to provide fully optimal answers to most of these questions, except for tests too short to be of much practical interest. So little is known about the answers to these questions, however, that we can learn much simply by trying out various plausible procedures and examining the kind of results obtained. This is the approach adopted here. Since convincing field studies (see Linn, Rock, and Cleary, 1969) of all the various plausible procedures would be impossibly extensive and expensive at this point, the procedures are here examined by test theory methods rather than by actual experimental test administrations.

In order to make any progress, we must somehow be able to predict how an examinee will perform on a new item, even if this item is at a different difficulty level from any to which the examinee has previously responded. To do this, we need to make use of item characteristic curves (see Birnbaum, 1968). For simplicity, it is assumed here that the available items differ from each other in difficulty but not on other statistical parameters.

Item Characteristic Curves

The characteristic curve of an item represents the probability of success on the item as a function of the ability level, θ , of the ex-

aminee. Here, we will consider only the case where the item characteristic curves are all of the form

$$P == P(\theta, b) == c + (1 - c)\Phi[a(\theta - b)], \quad (-\infty < \theta < \infty), \quad (1)$$

where the symbol $==$ is used to indicate a definition; a , b , c are parameters describing the item; and Φ represents the normal ogive function

$$\Phi(y) == \int_{-\infty}^y \phi(z) dz, \quad (2)$$

where ϕ is the normal curve ordinate

$$\phi(y) == \frac{1}{\sqrt{2\pi}} e^{-y^2/2}. \quad (3)$$

Several such curves are shown in Figure 1.

When the item cannot be answered correctly by random guessing, $c = 0$, and (1) becomes the usual normal ogive curve; the parameter c

$P_i(\theta)$

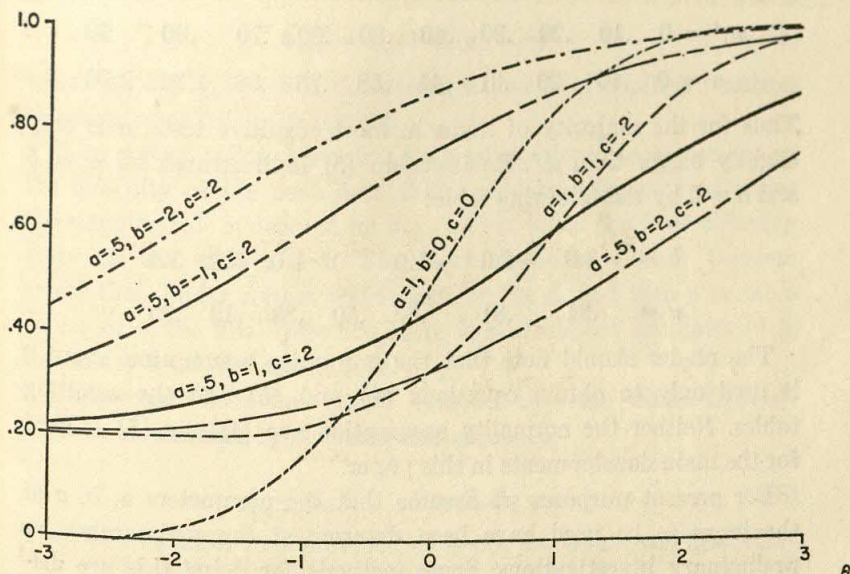


Figure 1. Normal ogive item characteristic curves.

culty score for large n is approximately normally distributed with variance of order $1/n$ and expectation $\theta^* = \theta + \text{constant}$. In the present problem,

$$\theta^* = \theta + \frac{1}{a} \Phi^{-1} \left(\frac{1 - \gamma}{1 - c} \right). \quad (9)$$

If $\gamma = \frac{1}{2}(1 + c)$, then $\theta^* = \theta$, the final difficulty score is a consistent estimator for θ , and its approximate large-sample variance is minimized by choosing

$$d_1 = \frac{\sqrt{2\pi}}{a(1 - c)}, \quad (10)$$

in which case this variance is

$$\text{Var } b_{n+1} = \frac{\pi(1 + c)}{2na^2(1 - c)}. \quad (11)$$

In contrast to the above, the results to be presented in this paper are small-sample results. The foregoing was presented here as background for the decision to limit investigation of shrinking-step-size procedures mainly to Robbins-Monro processes satisfying (8), with b_{n+1} recorded as the examinee's score.

Evaluation

In many problems of statistical inference, we require a consistent estimator, preferably one with the smallest possible sampling variance. Although estimating the examinee's ability level is definitely a statistical inference problem, neither of these requirements is appropriate here.

Equation (9) tells us that if $\gamma \neq \frac{1}{2}(1 + c)$, then b_{n+1} is not a consistent estimator of θ . Instead, it may be a strongly biased estimator, even when n becomes indefinitely large. However, the bias of b_{n+1} will be the same for all examinees tested. A constant bias does not affect the *relative* standing of examinees, and this is usually all that is important in mental testing. Thus, there is usually no need for us to seek either a consistent or an unbiased estimator of θ . (As a consequence, the present investigation need not be restricted to any particular value of γ .)

Actually, in most measurement situations the scale chosen for measuring θ is quite arbitrary. If θ in (1) is replaced by $\theta^* = \theta^2$ or by $\theta^* = \sqrt{\theta}$, we have a new three-parameter family of item char-

acteristic curves with ability now measured along a θ^* scale. In general, there is no good reason to assert that the θ scale provides a "truer" measure of ability than does the θ^* scale. Because of this fact, *comparisons between methods for estimating examinee ability must be invariant under any monotonic transformation of the scale used to measure ability.*

Here we will describe the effectiveness of any score x for measuring the ability θ by means of the information function of the scoring formula

$$I_x(\theta) = \frac{\left[\frac{\partial}{\partial \theta} \mathcal{E}(x | \theta) \right]^2}{\text{Var}(x | \theta)} \quad (12)$$

(Birnbaum, 1968, p. 453). The numerator here is the squared slope of the regression of test score on θ ; the denominator is the conditional variance of the test score for fixed θ .

The efficiency of scoring method x_2 relative to scoring method x_1 is given by the ratio

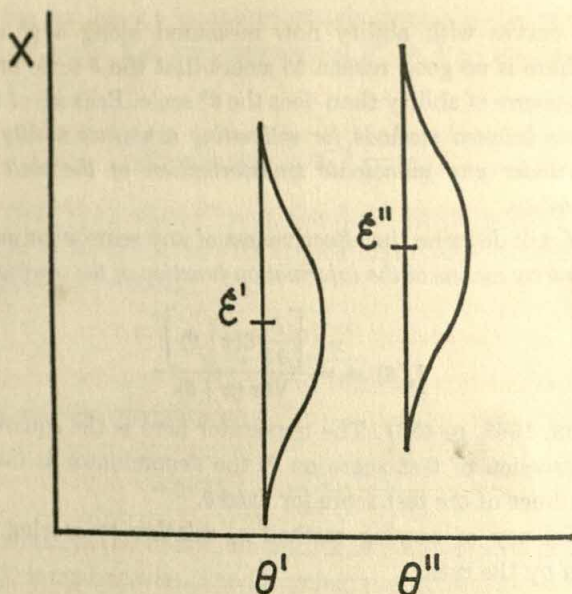
$$RE = \frac{I_{x_2}(\theta)}{I_{x_1}(\theta)}. \quad (13)$$

It is easily verified from (12) that the relative efficiency of two methods remains the same whether ability is measured by θ or by any monotonic transformation $\theta^*(\theta)$. This is the required invariance property.

Since we will not be assuming large n , we will not consider here the asymptotic properties of $I_x(\theta)$. The meaning and justification of $I_x(\theta)$ as used here will be described by paraphrasing Mandel and Stiehler (1954):

If it is desired to differentiate between two nearby values, θ' and θ'' , by means of the corresponding measurements x' and x'' , it is apparent that the success of the operation will depend on two circumstances: (1) the magnitude of the difference $\mathcal{E}'' - \mathcal{E}' = \mathcal{E}(x'' | \theta'') - \mathcal{E}(x' | \theta')$ for a given difference $\theta'' - \theta'$, i.e., the magnitude of the slope $(\mathcal{E}'' - \mathcal{E}')/(\theta'' - \theta')$; and (2) the precision of measurement $\text{Var}(x | \theta)$. These two desiderata can be combined in a single criterion, $I_x(\theta)$, defined as the ratio of the squared slope to $\text{Var}(x | \theta)$.

It is helpful to visualize the situation with the aid of a small diagram: A more formal discussion of the small-sample interpretation is given



by Lord (1952, pp. 21-25). (The term "information function" may be misleading if $I_x(\theta)$ is used without its asymptotic properties. It does not seem wise to try to rename $I_x(\theta)$ here, however.)

In order to have some idea of the meaning of any particular value of $I_x(\theta)$, it is helpful to compare it with values of $I_x(\theta)$ characterizing some "standard" test. The *standard tests* used here for such comparisons are n item conventional tests, composed of statistically equivalent items, administered in the usual way, the examinee's score being the number of items he answers correctly. Tailored tests on which there is no guessing are compared to a standard test on which $c = 0$; tailored tests with $c = .20$ are compared to a standard test with $c = .20$.

It may help to note that for a group in which θ is normally distributed with zero mean and unit variance, the 60-item standard test with $a = .50$, $b = 0$, and $c = 0$ will have a parallel-forms reliability of about .90. For tests like the standard tests, $I_x(\theta)$ is found from (12) to equal

$$I(\theta) = \frac{nP'^2}{PQ}, \quad (14)$$

where P is given by (1), P' is its derivative with respect to θ , and

$Q = 1 - P$. The reader may wish to look at Figures 2 and 3: in each, the tallest curve shows $I_x(\theta)/a^2$ for an appropriate standard test.

For the standard tests, $I_x(\theta)$ and $I(\theta)$ are proportional to test length, n . It is very helpful for understanding the meaning of $I_x(\theta)$ to think of a k fold increase in $I_x(\theta)$, for given θ , as equivalent to the increase in information gained by a k fold lengthening of a conventional test.

Computing the Information Function

Let us assume that b_1 , the difficulty of the first item administered, is the same for every examinee. When not otherwise stated, we will set $b_1 = 0$. Under Robbins-Monro procedures, n subsequent item difficulty parameters are determined from the examinee's responses in accordance with (7). The constants γ and d_1, d_2, \dots, d_n in (7) are selected in advance of the testing.

There are 2^n different possible patterns of item response $\mathbf{u} = \{u_1, u_2, \dots, u_n\}$. The probability of any particular pattern is simply

$$g(\mathbf{u} | \theta) = \prod_{i=1}^n P_i^{u_i} Q_i^{1-u_i}, \quad (15)$$

where P_i is the probability of success on the i th item administered, as determined from b_i and θ by equation (1). Since each pattern \mathbf{u} uniquely determines a final difficulty score $x = b_{n+1}$, (15) readily provides the conditional frequency distribution of final difficulty score for given θ , denoted by $f(x|\theta)$.

If n is 10 or 12, all 2^n values of (15) can be calculated by a computer for any chosen value of θ . This gives $f(x|\theta)$, from which $E(x|\theta)$ and $\text{Var}(x|\theta)$ are readily computed. This is repeated for various values of θ . The numerator of (12) can be computed by a recursion relationship, which need not be written out here.

When more than 12 or so items are administered, this brute-force method cannot be used. The results reported here were obtained by minor modifications of an ingenious method of Cochran and Davis (1965, p. 32), involving n successive Lagrangian interpolations. The numerator of (12) was computed for desired values of θ by evaluating the derivative of the appropriate Lagrangian interpolating polynomial.

Computations were done by a computer program devised by Martha Stocking. Final results were checked against brute-force results for small n and, in certain cases, against earlier results (Lord, 1971) for large n . Accuracy of repetitive interpolation, also checked by reruns with different interval widths, was found to be excellent.

*Choice of Initial Step Size When There
Is No Random Guessing*

In tailored testing, item difficulty is adjusted by successive steps in an attempt to match the item difficulty to the ability level of the examinee. If the steps are too small, it may take too long to reach a difficulty appropriate for a high-ability or low-ability examinee. If the step are too large, no close approximation to the appropriate difficulty level can be maintained, even if once achieved. The big theoretical advantage of a shrinking-step-size procedure is that it uses a large step size at first when the items may be poorly matched to the examinee's ability, and progressively smaller step sizes as the match is improved, avoiding both the above-mentioned problems.

In (10), an asymptotically "optimum" value of $d_1 = \sqrt{2\pi}/a$ was mentioned for $c = 0$. However, this value was chosen to minimize the asymptotic sampling variance of the final difficulty score, without reference to the information function $I_x(\theta)$ that is of concern here.

Figure 2 shows a plot of the information function for each of five initial step sizes, determined by $d_1 = 1/a, 2/a, \sqrt{2\pi}/a, 3/a$, and $5/a$. The stepping procedure is Robbins-Monro as defined by (7) and (8). The score is the final difficulty score $x = b_{n+1}$. All five have $n = 60, c = 0$, and $\gamma = .5$.

The figure is labeled so that it can be used for any choice of initial item difficulty b_1 and for any value of the item discriminating-power parameter a . This is possible because (1) remains unchanged and (7) and (12) remain valid if we add a constant to θ and subtract the same constant from b ; also if we multiply a by some constant k , divide θ and each b by k , and multiply $I_x(\theta)$ by k^2 . This is the reason why the horizontal axis is labeled $a(\theta - b_1)$, not just θ , and why the vertical axis is labeled $I_x(\theta)/a^2$. Typical test items might have values of $a = .5$ approximately.

We do *not* assume anything here about the frequency distribution of ability. When we use Figure 2, however, we need some idea as to where the examinees fall along the horizontal axis. For

example, if we are chiefly interested in examinees in the range $-2 \leq \theta \leq 3$, then we need look at $I_x(\theta)/a^2$ in Figure 2 only for those portions of the curves in the range $a(-2 - b_1) \leq a(\theta - b_1)$

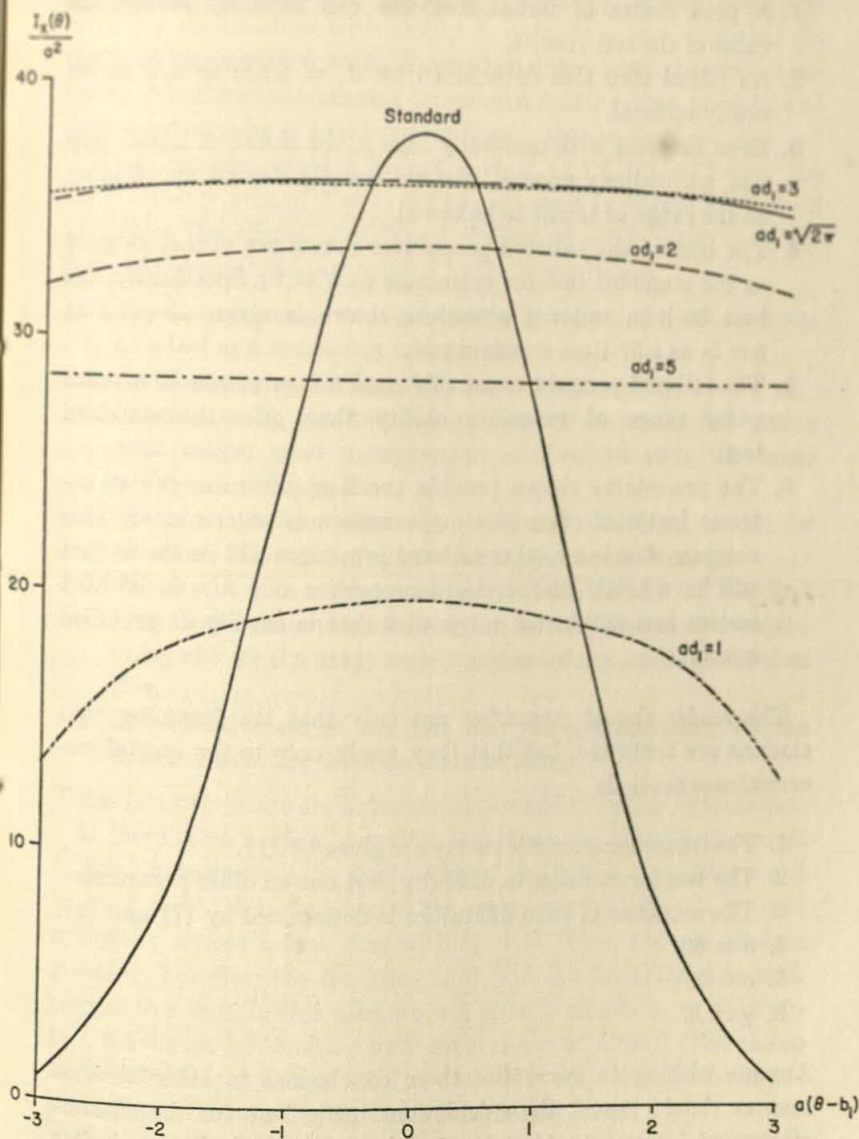


Figure 2. Index of discrimination for five Robbins-Monro tailored testing procedures with $n = 60$, $c = 0$, and $\gamma = .5$.

$\leq a(3 - b_1)$. If $a = .5$ and $b_1 = 0$, we will want only the range $-2 \leq \theta \leq 1.3$.

Some of the conclusions suggested by Figure 2 are:

1. A poor choice of initial step size can seriously reduce the value of the test results.
2. An initial step size determined by $d_1 = 2.5/a$ or $3/a$ seems nearly optimal.
3. Even for tests with unusually high a , the choice of initial step size, surprisingly enough, will not usually depend appreciably on the range of talent to be tested.
4. The best of the tailored procedures shown are almost as good as the standard test for examinees at $\theta = b_1$. Specifically, the best 60-item tailored procedure shown is about as good at $\theta = b_1$ as a 57-item standard test.
5. The tailored procedures provide good measurement for a much wider range of examinee ability than does the standard test.
6. The procedures shown provide excellent measurement at extreme levels of $a\theta$, where measurement is *not* required. This suggests that some other tailored procedure will be found that will have better measurement properties near $\theta = b_1$ without serious loss within the range of θ that is usually of practical interest.

The reader should remember not only that the foregoing conclusions are tentative, but that they apply only to the special circumstances studied:

1. The item characteristic curves are given by (1).
2. The test items differ in difficulty, but not on other parameters.
3. The sequence of item difficulties is determined by (7) and (8).
4. $n = 60$.
5. $c = 0$.
6. $\gamma = .5$.

Anyone wishing to generalize these conclusions to other circumstances should repeat the calculations made here for the circumstances of interest to him. It must be understood that a similar warning applies to conclusions drawn in the sections that follow.

Choice of Offset and of Initial Item Difficulty

The parameter γ in (7) will be called the *offset*. It determines the relative size of upward and downward steps. When items cannot be answered correctly by random guessing, optimal measurement for an examinee with ability θ_0 requires a test composed entirely of items with $b = \theta_0$. Thus when there is no guessing, symmetry considerations require in tailored testing that upward and downward steps be of equal size for fixed i ; that is, that $\gamma = \frac{1}{2}$.

When random guessing occurs, $c \neq 0$, and $I_x(\theta)$ no longer has the symmetry around $\theta = b_1$ that is apparent in Figure 2. Figure 3 illustrates the effects of different choices of offset when $c = .20$. Some of the conclusions suggested by these results are

1. Random guessing may substantially reduce the effectiveness of a test as a measuring instrument.
2. An offset of $\gamma = .5$ is unsatisfactory when $c = .20$.
3. When $\gamma = \frac{2}{3}$, the information function is nearly flat over a wide range: good measurement is obtained over the range $-1 \leq a(\theta - b_1) \leq 4$.
4. When γ is reduced from $\frac{2}{3}$ to $.6$, more information can be obtained, although for a narrower range of examinees.
5. Where there is random guessing, the difficulty of the first item should ordinarily not be set at $b_1 = 0$, that is, at the mean value of θ for the group tested; instead it should be easier than this.
6. The effectiveness of the test may depend markedly on the choice of difficulty level for the first item.

These last two points are important and require further explanation.

If there is no random guessing, and if we are equally interested in good measurement on both sides of the mean (i.e., above and below $\theta = 0$), then, because of the symmetry in Figure 2, we will ordinarily choose a first item with $b_1 = 0$. When there is random guessing, however, the function $I_x(\theta)$ will not be symmetric with respect to $\theta = 0$. In this case, we will usually choose b_1 so as to obtain maximum information over some range of ability. The choice of $b_1 \neq 0$ does not change the $I_x(\theta)$ curve; it only changes the interpretation of the base line on which it is plotted.

For example, Figure 3 shows that when $\gamma = \frac{2}{3}$ we will have good discrimination over the range $-1 \leq a(\theta - b_1) \leq 4$. If

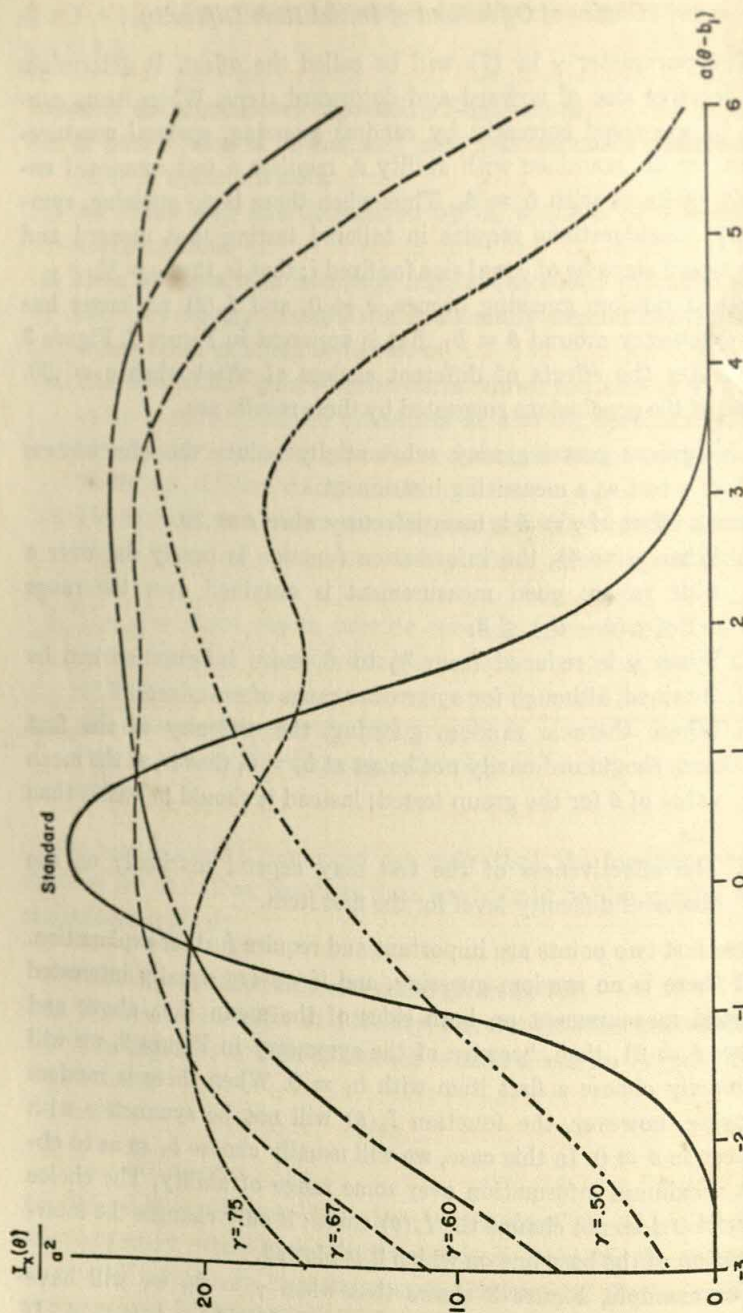


Figure 3. Index of discrimination for four Robbins-Monro procedures with $n = 60$, $c = .20$, and $d_s = 3.1/a$.

$a = 1$, the choice of $b_1 = -1$ makes this range $-1 \leq \theta + 1 \leq 4$ or $-2 \leq \theta \leq 3$.

Figure 3 shows that still more information than this can be obtained by setting $\gamma = .6$. In this case, best results are obtained for an interval such as $1 \leq a(\theta - b_1) \leq 4$. If $a = \frac{1}{2}$, the choice of $b_1 = -5$ will give good measurement in the range $1 \leq \frac{1}{2}\theta + \frac{5}{2} \leq 4$ or $-3 \leq \theta \leq 3$.

At first sight, this last result seems so strange as to require some rationalization. It shows that if $a = .5$, best results are obtained by setting $\gamma = .6$ and making the first item so easy ($b_1 = -5$) that virtually everyone will answer it correctly! The following table shows $P(\theta, -5)$, the probability of answering this item correctly, for different ability levels:

$\theta =$	-3	-2	-1	0
$P(\theta, -5) =$.87	.95	.98	.995

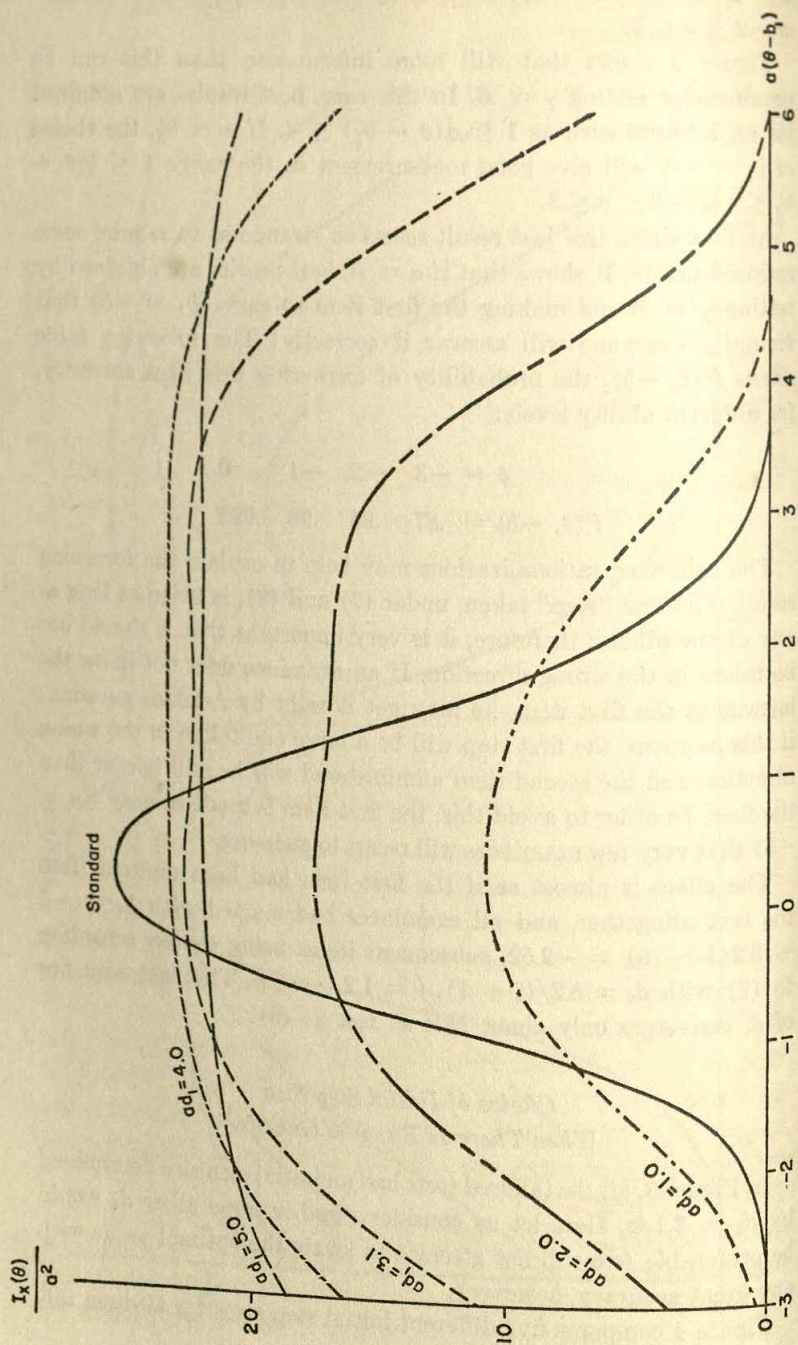
The following rationalizations may help to explain the foregoing result. The first "step" taken, under (7) and (8), is twice as long as any of the others; therefore, it is very important that it should not be taken in the wrong direction. If an examinee does not know the answer to the first item, he may get it right by random guessing; if this happens, the first step will be a large one taken in the wrong direction and the second item administered will be still harder than the first. In order to avoid this, the first item is made so easy ($b_1 = -5$) that very few examinees will resort to guessing.

The effect is almost as if the first item had been omitted from the test altogether, and all examinees had started with $b_1 = -5 + 6.2(1 - .6) = -2.52$, subsequent items being chosen according to (7) with $d_i = 6.2/(i + 1)$, $i = 1, 2, \dots, n$. This last sequence of d_i converges only about half as fast as (8).

Choice of Initial Step Size When There Is Random Guessing

In Figure 3, all the tailored tests had an initial step size determined by $d_1 = 3.1/a$. Here let us consider whether some other d_1 would be preferable (we will not attempt to locate the optimal value with any great accuracy, however).

Figure 4 compares five different initial step sizes for 60-item tai-



lored tests having $c = .20$ and $\gamma = 2/3$. This figure suggests the following conclusions:

1. If the length of the first step is badly chosen, measurement may be seriously impaired.
2. Good measurement is obtained if d_1 is in the range $3 \leq ad_1 \leq 5$.
3. The larger the initial step size, the larger the range of effective measurement.
4. The larger the initial step size, the easier the first item should be.
5. If $a = .5$, a choice of $d_1 = 8.0$ and of $b_1 = -3$ gives good measurement when $-1 \leq .5(\theta + 3) \leq 4$, that is, for $-5 \leq \theta \leq 5$.
6. The foregoing 60-item tailored test is about as effective for all examinees in the range $-3 \leq \theta \leq 3$ as a 55-item standard test is for those examinees it measures best.

When $\gamma = .6$, the curves for the information function (not shown here) tend to be less flat and skewed to the left. The curve for $ad_1 = 3.1$ and $\gamma = .6$ is barely higher in the range $2.5 \leq \theta \leq 3.5$ than the curve for $ad_1 = 4.0$ and $\gamma = 2/3$, but it is not as high elsewhere.

Fixed vs. Shrinking Step Size

Lord (1971) investigated fixed-step-size procedures, item difficulty being determined by (6). When there is no guessing, the "best" 60-item fixed-step-size procedure found uses a step size of $d = .20/a$. The information curve for this procedure is shown in Figure 5, labeled $ad = .20$. Other curves are shown, for a fixed step size of $ad = .05$ and for $ad = .50$.

One of the best curves from Figure 2 is copied in Figure 5 for comparison. This is the curve for the Robbins-Monro procedure with $ad_1 = \sqrt{2\pi}$, so labeled here. It seems clear that in this case, at least, the Robbins-Monro procedure is better than any of the fixed-step-size procedures studied.

When step size is fixed, use of final difficulty score b_{n+1} is often unsatisfactory: when step size is fixed, such a score can assume only a particular set of more or less widely spaced values and thus cannot approximate θ as closely as might be desired. It is usually better to

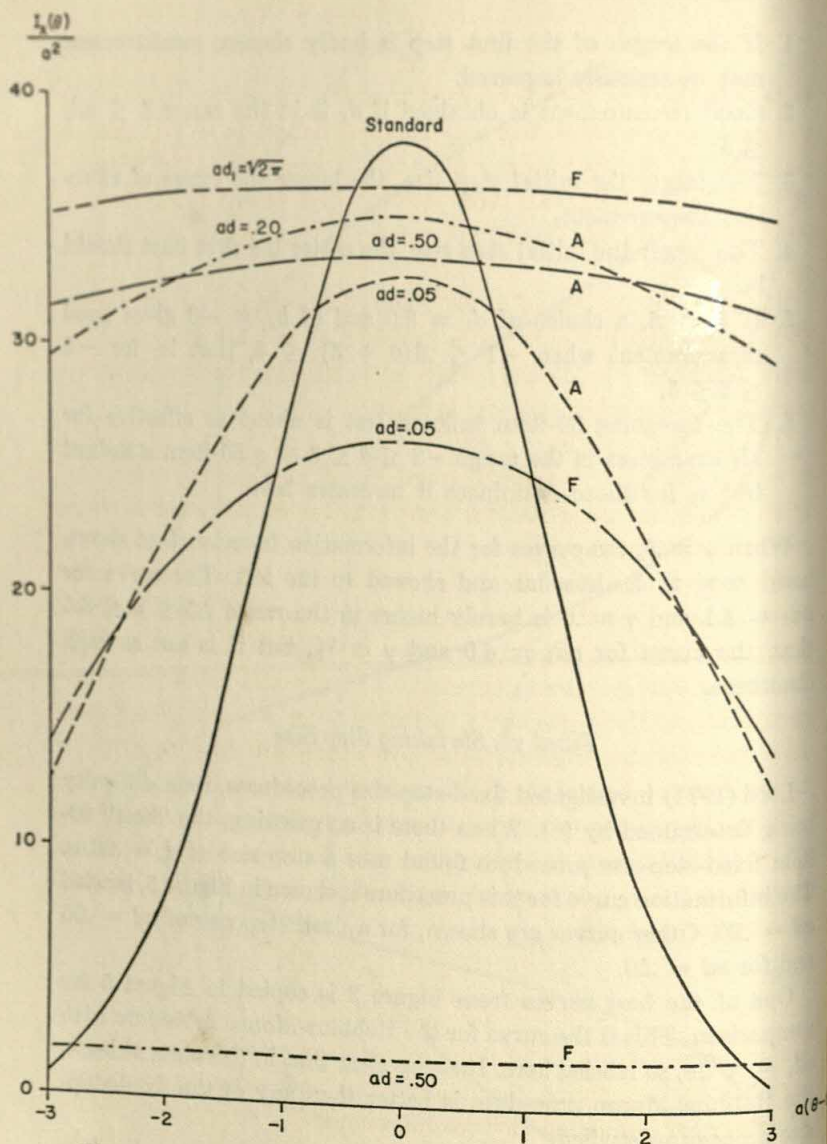


Figure 5. Comparison of fixed-step-size procedures with one of the best Robbins-Monro procedures. $n = 60$, $c = 0$, $\gamma = .50$ (A = average difficulty score, F = final difficulty score).

score the test by adding up (or averaging) b_i values. The result is called the *average difficulty score*, defined precisely by

$$x = \frac{1}{n} \sum_{i=2}^{n+1} b_i. \quad (16)$$

Note that the first item is omitted from the summation (since it is the same for everyone), but that the hypothetical $(n+1)$ st item is included. [A slightly superior method of scoring found by Wetherill, Chen, and Vasudeva (1966) has not been investigated for tailored testing. Its use should not materially change the conclusions reached here.]

For $ad = .05$ and for $ad = .50$, Figure 5 shows information curves both for average difficulty score (marked A) and for final difficulty score (marked F). Clearly, when $n = 60$, the average difficulty score is superior when step size is fixed.

The conclusions already drawn from Figure 5 may be typical of results when $c = 0$ and n is large. A look at Figure 6, which displays results for some 10-item tests, suggests caution against overgeneralization, however.

When $n = 10$, a step size determined by $ad_1 = 3$ was found to be "best" among the Robbins-Monro procedures tried. The information curve for this procedure is shown in Figure 6, labeled $ad_1 = 3$. The other curves shown represent fixed-step-size procedures corresponding to those displayed in Figure 5 for $n = 60$. For θ near zero, some of the fixed-step-size procedures are here seen to be better than the best of the Robbins-Monro procedures.

It would be good to have some results evaluating the use of average difficulty score in conjunction with Robbins-Monro stepping procedures. Also for various weighted average difficulty scores, such as, for example,

$$x = \frac{2}{n(n+3)} \sum_{i=2}^{n+1} ib_i.$$

Unfortunately, for $n > 10$ there do not seem to be any convenient methods for computing the expectations and sampling variances necessary for evaluating such scoring methods.

Item Economy

In planning tailored testing, one must have items available for each of the possible "paths" that the examinee may follow as a re-

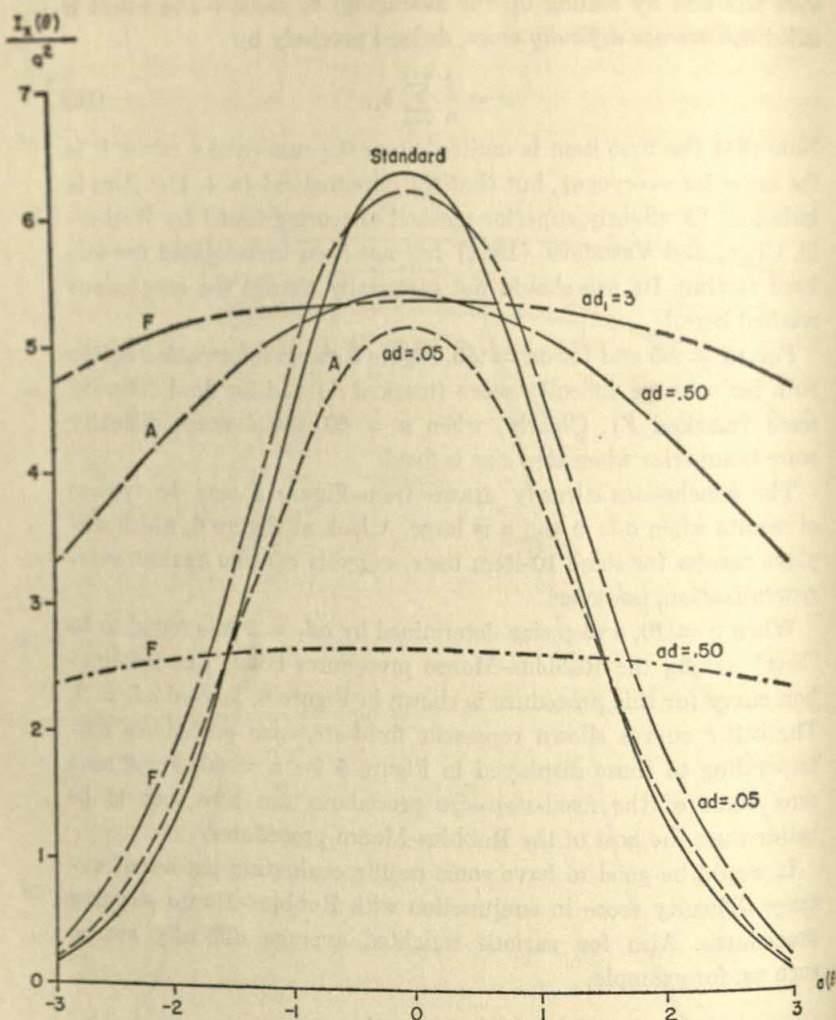


Figure 6. Comparison of fixed-step-size procedures with one of the best Robbins-Monro procedures. $n = 10$, $c = 0$, $\gamma = .50$ (A = average difficulty score, F = final difficulty score).

sult of the stepping procedure. In the case of Robbins-Monro procedures with step size governed by (8), one needs to have two different items available for administration following the first item, four additional items available for administration following the second item, and so forth. A total of $2^n - 1$ items are needed, in

theory, before testing can be begun. Even for $n = 20$, this would be more than a million items.

For such shrinking-step-size procedures, the number of items needed is represented by a geometric series in powers of 2 with n terms. For fixed-step-size procedures, however, only an arithmetic series is involved, so that only $n(n+1)/2$ items are required. For $n = 20$, 210 items are needed in theory when step size is fixed; for $n = 60$, 1830 items. This number can be greatly reduced by certain obvious shortcuts and approximations.

These figures suggest that for $n > 6$, say, a strict use of Robbins-Monro methods is impractical because of the number of items required. An obvious suggestion is to approximate the steps required by (8), while using only steps that are a multiple of some prechosen minimum step size, denoted here by Δ . If Δ is large enough, the number of items needed is much reduced.

For the case where there is no guessing, Figure 7 compares information curves obtained for two different values of Δ with those obtained for other procedures already studied. The top curve is the same Robbins-Monro process used as a standard of comparison in Figure 5. The curve just below it is the same process, modified with $\Delta = .05/a$. The bottom curve, labeled $ad_1 = 2.6$, $a\Delta = .20$, is also the same process, modified with $\Delta = .20/a$. The values of ad_1 are compared in the following table for these three curves:

	$i = 1$	2	3	10	30	60
eq. (8)	2.51	1.25	0.84 ...	0.25 ...	0.08 ...	0.04
$\Delta = .05/a$	2.5	1.25	0.85 ...	0.25 ...	0.10 ...	0.05
$\Delta = .20/a$	2.6	1.2	0.80 ...	0.20 ...	0.20 ...	0.20

For comparison, Figure 7 repeats some of the curves from Figure 5, representing fixed-step-size procedures. The figure suggests the following tentative conclusions.

1. When Δ is not much larger than d_1/n , modification of the Robbins-Monro procedure does not cost much in terms of measurement efficiency. However, this is the case where the modification gains little in item economy.
2. When Δ is enough larger than d_1/n to economize effectively on items, the suggested modification of the Robbins-Monro process causes an unacceptable penalty on the efficiency of measurement.

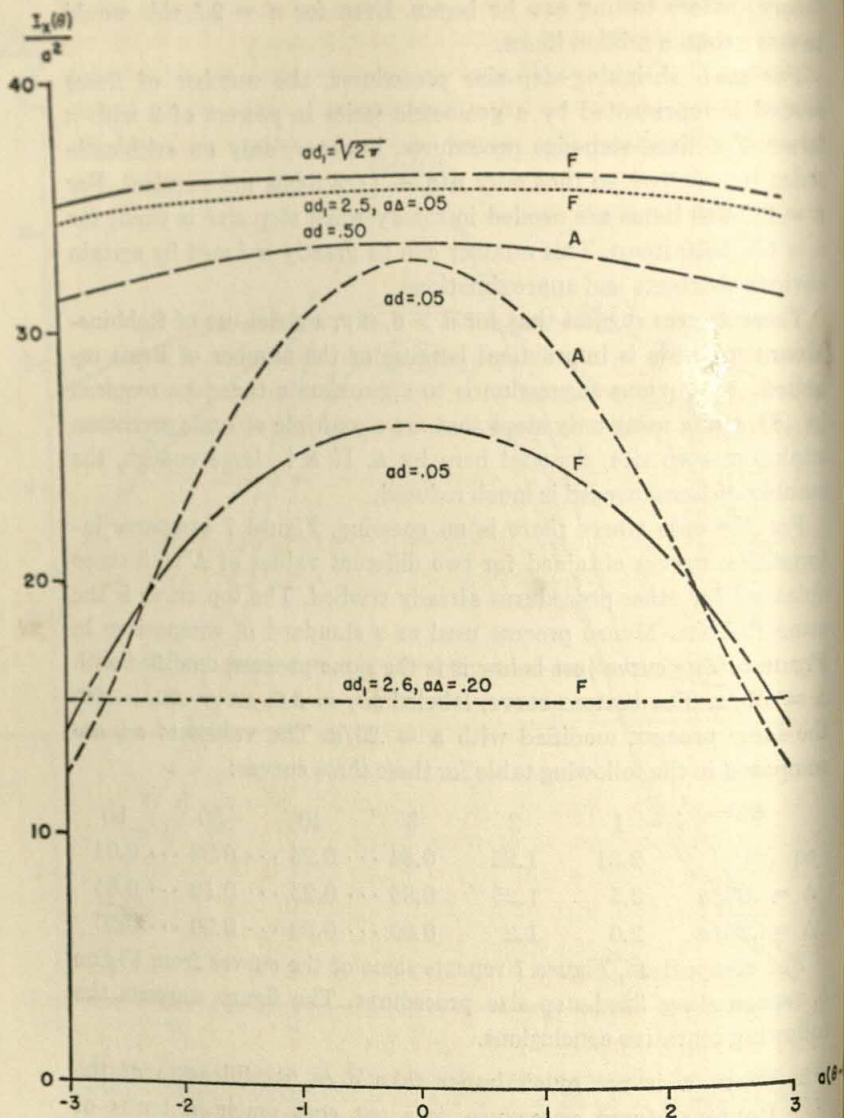


Figure 7. Comparison of modified and unmodified Robbins-Monro procedures, also fixed-step-size procedures, $n = 60$, $c = 0$, $\gamma = .50$.

Actually, the stepping procedure used when $a\Delta = .20$ might seem to have advantages over a comparable fixed-step-size procedure. The reason is that the large first steps allow quick correction of any initial misjudgment of the item difficulty level appropriate

for the examinee. Why, then, does this procedure show up so poorly? The answer may be that the poor results are due to the use of final difficulty score. If average difficulty score were used with this stepping procedure, better measurement might result. Unfortunately, this possibility cannot be checked here, because of the difficulty in computing information curves for this scoring procedure when the d_i are unequal.

Hybrid Procedures

Several writers have suggested use of fixed-step-size procedures preceded by one or more large initial steps. This seems a promising approach, for reasons mentioned in the last section. A computer program was written to investigate the following hybrid procedure:

1. First take n_1 steps of decreasing size, as determined by (7) and (8);
2. then take $n - n_1$ steps of fixed size, d , determined by (6).
3. Compute an average difficulty score from the last $n - n_1$ items only.

For the case where there is no guessing, Figure 8 shows the effectiveness of measurement obtained by three such hybrid procedures, together with results for two fixed-step-size procedures, for comparison.

In general, the effect of the hybridization here seems to be to improve measurement at extreme values of θ at the expense of measurement around $\theta = 0$. If $a = .5$ and $b_1 = 0$, the simple fixed-step-size procedure with $ad = .20$ is better than any of the hybrid procedures investigated within the range $-3 \leq \theta \leq 3$.

The curve for the hybrid procedure with $ad_1 = 1$, $n_1 = 1$, and $ad = .40$ is not shown in the figure because it almost entirely coincides with the fixed-step-size curve for $ad = .40$. Thus for this fixed step size, such hybridization is neither helpful nor harmful. The results are still inferior for most purposes, however, to those obtained with a smaller fixed step size.

Figure 9 shows results obtained for various hybrid methods when random guessing occurs. Here, there is little to choose between certain of the hybrid procedures and the comparable pure fixed-step-size procedure. Hybridization seems to offer no distinct advantages, however.

Many other hybrid procedures would be of interest (see Wetherill, 1963). No others are considered here.

Summary and Conclusions

An earlier study (Lord, 1971) investigated various fixed-step-size methods for tailored testing. It seems plausible that shrinking-step-size methods might be preferable. These allow *rapid* matching of item difficulty to the ability level of the examinee initially, when his level is very poorly known; but *close* matching later in the testing, when his level can be inferred with some accuracy from responses to items already administered.

Robbins-Monro procedures are shrinking-step-size procedures in which the final score of the examinee (called the final difficulty score) is the difficulty level of the item that would be administered next if the testing were continued. Other scoring methods would no doubt be preferable for certain of these procedures, but computational difficulties have prevented any extensive investigation of them here.

The Robbins-Monro procedures studied here have a harmonic sequence of step lengths, or at least an approximation to this. There is no firm basis for this choice; other sequences also deserve investigation.

Tailored testing methods involve many different parameters:

1. examinee ability level (θ)
2. test length (n)
3. item difficulty ($b_i, i = 1, \dots, n$)
4. item discriminating power ($a_i, i = 1, \dots, n$)
5. item guessing parameters ($c_i, i = 1, \dots, n$)
6. offset (γ)
7. parameters controlling step size ($d_i, i = 1, \dots, n$).

In addition, there is a virtually unlimited choice of scoring methods. With so many variables to control, it is difficult to reach any really firm conclusions about the general merits of various procedures on the basis of a few illustrative specimens of each. The following list includes tentative conclusions at various levels of generality. The reader is cautioned that the conclusions are based on very limited investigations; many of them cannot be expected to hold for other circumstances.

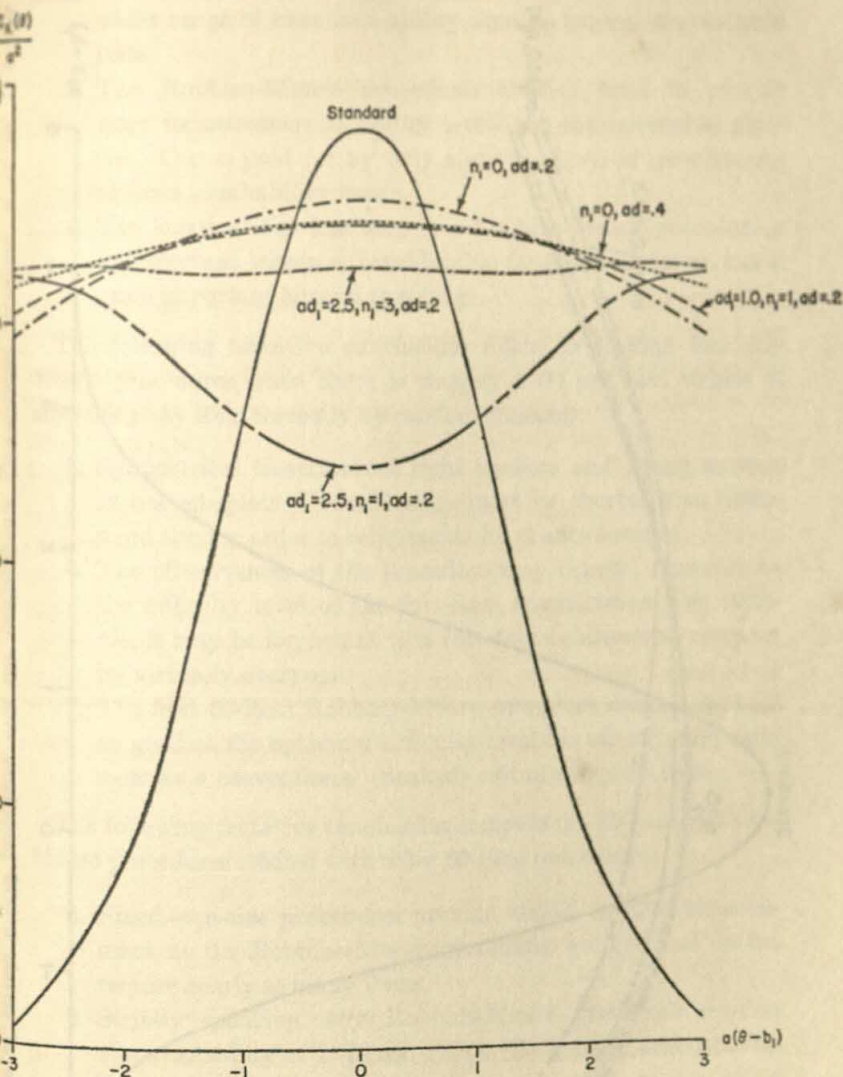


Figure 8. Comparison of certain hybrid procedures with the corresponding fixed-step-size procedures, $n = 60$, $c = 0$, $\gamma = .50$.

1. When there is no random guessing, the best 60-item Robbins-Monro procedure studied is about as good at the optimal difficulty level for effective measurement as a conventional (peaked) test of about 57 items.

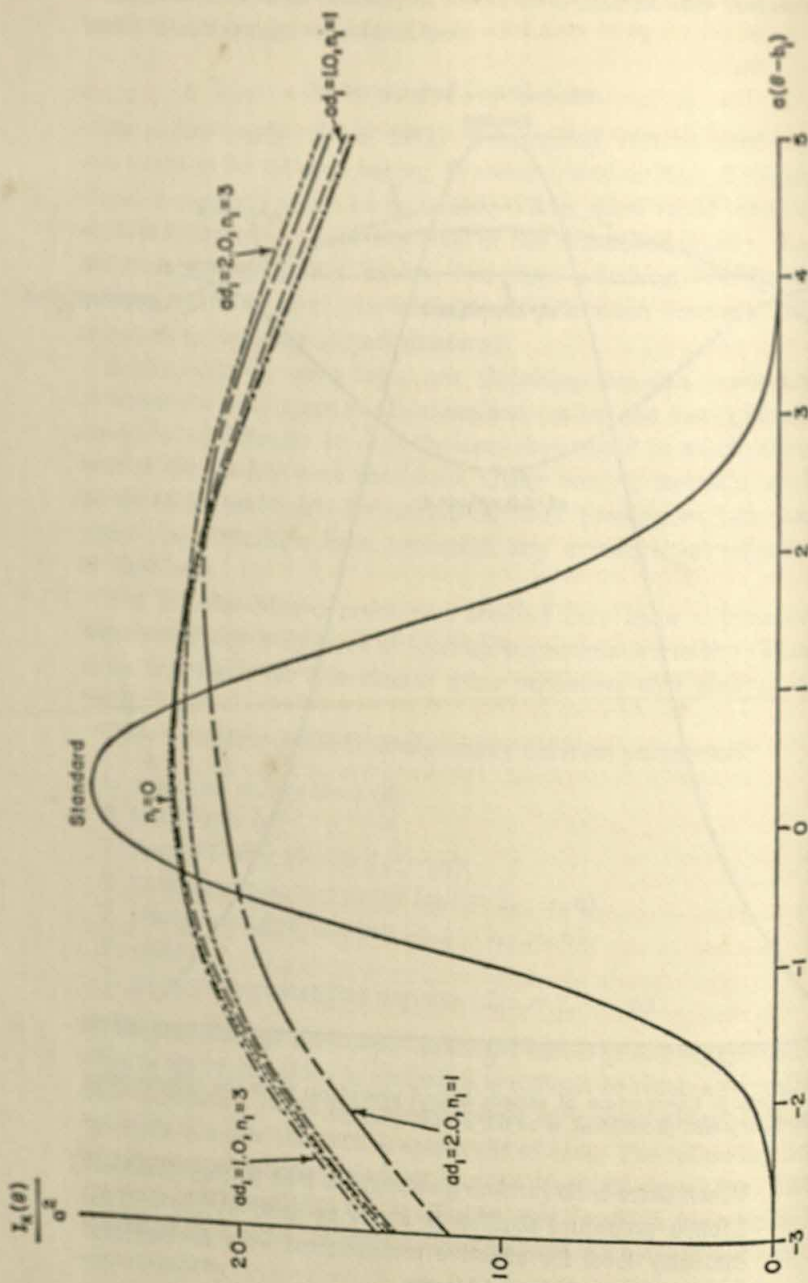


Figure 9. Comparison of normal, lognormal, and beta distributions.

2. Tailored procedures provide good measurement over a much wider range of examinee ability than do typical conventional tests.
3. The Robbins-Monro procedures studied tend to provide good measurement at ability levels not encountered in practice. This is paid for by only slightly impaired measurement at more usual ability levels.
4. The length of the first step in a Robbins-Monro procedure is unimportant within a considerable range of tolerance, but is quite important beyond this range.

The following tentative conclusions relate to 60-item Robbins-Monro procedures when there is roughly a 20 per cent chance of answering any item correctly by random guessing:

5. Symmetrical treatment of right answers and wrong answers is not adequate. Upward steps must be shorter than downward steps in order to compensate for chance success.
6. The effectiveness of the procedure may depend markedly on the difficulty level of the first item administered. For example, it may be important that this item be answered correctly by virtually everyone.
7. The best 60-item Robbins-Monro procedure studied is about as good at the optimum difficulty level for effective measurement as a conventional (peaked) test of about 55 items.

The following tentative conclusions compare the 60-item Robbins-Monro procedures studied with other 60-item procedures:

8. Fixed-step-size procedures provide almost as good measurement as the Robbins-Monro procedures studied and do not require nearly so many items.
9. Strictly speaking, any Robbins-Monro procedure requires the availability of 2^n items. This is likely to be uneconomical for $n > 6$, say.
10. One shortcut, used to reduce the number of items needed to reasonable limits, destroyed the measurement effectiveness of the Robbins-Monro procedure. It might be possible to regain most of this loss by changing the scoring method. Computational complications prevented further investigation of this possibility.

Certain hybrid procedures were investigated: a Robbins-Monro process with large step sizes was used for n_1 initial steps, after which a fixed-step-size procedure was used with smaller steps.

11. Hybridization improved measurement at extreme ability levels, but often at the cost of impaired measurement for levels usually encountered. For typical items, the hybrid procedures studied showed no advantage over fixed-step-size procedures for virtually all examinees in a typical group.

To summarize, shrinking-step-size procedures have certain obvious advantages over fixed-step-size procedures. However, if more than six or seven items are to be administered to an examinee, the item pool required by the shrinking-step-size procedures is so large as to be prohibitive. Certain obvious shortcuts are possible for reducing the item pool, but so far these do not seem to lead to as effective measurement as do the simple fixed-step-size procedures.

Only a few of the possible procedures and circumstances have been explored here. Many of the tentative conclusions stated here will not be valid for all circumstances. Other procedures and circumstances should be widely investigated. In particular, anyone designing a tailored test and preparing it for actual use should carry through on his own account investigations similar to those reported here, choosing the parameters and procedures studied to fit his own special requirements.

It should be noted that no method of administering items and scoring *dichotomous* item responses can produce better measurement than is achieved at $\theta = 0$ by one of the "standard tests." The tailored test tries to approach this level of measurement for all examinees, not just those at $\theta = 0$. Can such tests be further much improved, within the context of dichotomous item response?

REFERENCES

- Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968. Chapters 17-20.
- Bock, R. D. and Lieberman, M. Fitting a response model for n dichotomous items. Research Memorandum No. 8. Chicago: Department of Education, Statistical Laboratory, University of Chicago, 1967.
- Cochran, W. G. and Davis, M. The Robbins-Monro method for

- estimating the median lethal dose. *The Journal of the Royal Statistical Society, Series B*, 1965, 27, 28-44.
- Freeman, P. R. Optimal Bayesian sequential estimation of the median effective dose. *Biometrika*, 1970, 57, 79-89.
- Hodges, J. L., Jr. and Lehmann, E. L. Two approximations to the Robbins-Monro process. In J. Neyman (Ed.), *Proceedings of the third Berkeley symposium on mathematical statistics and probability*. Volume 1. Berkeley: University of California Press, 1956.
- Linn, R. L., Rock, D. A., and Cleary, T. A. The development and evaluation of several programmed testing methods. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 129-146.
- Lord, F. M. A theory of test scores. *Psychometric Monograph*, 1952, No. 7.
- Lord, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 989-1020.
- Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer assisted instruction, testing, and guidance*. New York: Harper & Row, 1971.
- Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.
- Mandel, J. and Stiehler, R. D. Sensitivity—A criterion for the comparison of methods of test. *Journal of Research of the National Bureau of Standards*, 1954, 53, 155-159.
- Owen, R. J. A Bayesian approach to tailored testing. *Research Bulletin 69-92*. Princeton, N. J.: Educational Testing Service, 1969.
- Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951, 22, 400-407.
- Turnbull, W. W. Relevance in testing. *Science*, 1968, 160, 1424-1429.
- Wasan, M. T. *Stochastic approximation*. Cambridge: Cambridge University Press, 1969.
- Wetherill, G. B. Sequential estimation of quantal response curves. *Journal of the Royal Statistical Society*, 1963, 25, 1-38.
- Wetherill, G. B., Chen, H., and Vasudeva, R. B. Sequential estimation of quantal response curves: a new method of estimation. *Biometrika*, 1966, 53, 439-454.
- Wetherill, G. B. and Levitt, H. Sequential estimation of points on a psychometric function. *British Journal of Mathematical and Statistical Psychology*, 1965, 18, 1-10.

RELAXED RANK ORDER TYPAL ANALYSIS¹

LOUIS L. McQUITTY

University of Miami
Coral Gables, Florida

RANK Order Typal Analysis uses a strict definition for classifying objects into types; few objects satisfy its requirements. The method is relatively inappropriate for fallible data (McQuitty, 1963).

This paper relaxes the definition of types in relation to the requirements of data and renders Rank Order Typal Analysis more widely applicable for classifying objects into hierarchical systems based on assessment of their characteristics.

If the initial classification criterion is relatively restrictive, the method will usually initiate a classification but the classification may not proceed far. At this point, the criterion is liberalized minimally in order for the classification to proceed.

On the other hand, if the initial classification criterion is liberal for the data, the method will either classify every object into one of only two categories or even fail to render a division. In either of these two eventualities, the criterion can be made more restrictive. The classification criterion is adjusted to the requirements of the data as the analysis proceeds.

The Theory

In Rank Order Typal Analysis, a type is defined as a category of objects (specified in terms of selected characteristics) of such a nature that every object in the category is more like every other object in the category than it is like any object in any other category.

¹This investigation was supported by Public Health Service Research Grant No. MH 14070-01 from National Institute of Mental Health.

Furthermore, if only pure categories are sought a category of n objects does not qualify as a type unless it includes qualifying sub-categories of 2, 3, 4 \dots $n - 1$ objects, where n equals the number of objects in the category under consideration (McQuitty, 1964).

By way of contrast, Elementary Linkage Analysis (McQuitty, 1957) requires only that every object is classified with the one object most like itself. Between these two extremes are several degrees of freedom which can be used to adjust the method to the requirements of the data.

The Method

Rank Order Typal Analysis requires that the indices of a matrix of interassociations be converted to ranks within columns.

Whenever an object is introduced into a category it brings with it two groups of ranks, viz., those that it has in its column, showing the order in which other objects rank with it, and those that it has in its row, showing the ranks that it has with other objects.

In Relaxed Rank Order Typal Analysis, a type is a category of objects of such a nature that an object has no rank in either its row or column above a specified maximum; the maximum is the minimal value which renders an internally consistent category.

Illustration

The method is illustrated with the data of Table 1, which reports agreement scores between pictures of spoons as judged by one subject in terms of specified characteristics. These data were chosen because they had proven particularly difficult to analyze with the method of Hierarchical Classification by Reciprocal Pairs; an improved method had to be developed in order to classify the spoons (McQuitty, Price, and Clark, 1967).

The entries in every column were ranked from 1 to $n - 1$, where n equals the number of objects. Table 2 reports these ranks within columns. A slight deviation was introduced into the usual method of ranking. Suppose, for example, that the highest score is 30, followed by 29, 29, 29, and 28. Thirty is assigned a rank of 1, and every 29 a rank of 2 (rather than 3 as in the customary method); 28 is assigned a rank of 5 because it has four scores above it. The smallest numerical rank involved is assigned to all tied scores.

Reciprocal pairs. A search is made first for the "strict" types,

TABLE 1*

Agreement Scores between Objects

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		20	29	20	25	24	20	13	29	20	23	18	28	28	24	30	28	16	22	20
2	30		20	25	17	15	26	20	15	25	13	26	30	20	27	21	25	20	18	25
3	29	20		19	26	34	23	13	33	20	30	18	27	31	26	34	32	19	30	22
4	30	25	19		22	18	25	20	18	22	18	27	20	19	25	19	22	21	19	25
5	25	17	26	22		24	17	23	23	16	28	20	18	23	20	21	24	14	26	14
6	24	15	34	18	24		20	12	33	17	32	15	29	29	23	30	28	16	32	20
7	30	26	23	25	17	20		14	19	30	18	26	18	22	25	23	26	24	16	28
8	13	20	13	20	23	12	14		12	15	16	18	13	11	21	11	13	15	18	12
9	29	15	33	18	23	33	19	12		20	29	15	30	28	26	33	25	17	29	21
10	30	25	20	22	16	17	30	15	20		14	20	21	21	28	22	24	27	15	31
11	23	13	30	18	28	32	18	16	29	14		16	25	24	21	29	22	13	31	16
12	18	26	18	27	20	15	26	18	15	20	16		15	17	23	15	19	21	18	21
13	28	20	27	20	18	29	18	13	30	21	25	15		27	26	30	26	17	23	23
14	28	20	31	19	23	29	22	11	28	21	24	17	27		24	30	27	19	23	22
15	24	27	26	25	20	23	25	21	26	28	21	23	26	24		25	24	22	24	27
16	30	21	34	19	21	30	23	11	33	22	29	15	30	30	25		28	17	26	23
17	28	25	32	22	24	28	26	13	25	24	22	19	26	27	24	28		15	22	22
18	16	20	19	21	14	16	24	15	17	27	13	21	17	19	22	17	18		14	30
19	22	18	30	19	26	32	16	18	29	15	31	18	23	23	24	26	22	14		16
20	20	25	22	25	14	20	28	12	21	31	16	21	23	22	27	23	22	30	16	

* From McQuitty, Price, and Clark, 1967.

TABLE 2

Table 1 by Ranks within Columns

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		9	8	10	4	9	12	11	5	11	9	10	4	4	10	3	2	14	10	14
2	12		14	2	15	17	3	3	17	5	18	2	13	15	2	14	8	7	13	5
3	2	9		13	2	1	9	11	1	11	3	10	5	1	4	1	1	8	3	9
4	12	4	16		10	14	6	3	15	7	12	1	13	16	7	16	13	5	12	5
5	7	16	10	6		9	17	1	11	16	6	7	15	10	19	14	10	17	5	18
6	8	17	1	17	5		12	15	1	15	1	16	3	3	14	3	2	14	1	14
7	12	2	12	2	15	12		10	14	2	12	2	15	12	7	11	6	3	16	3
8	19	9	19	10	7	19	19		19	17	14	10	19	19	17	19	19	16	13	19
9	2	17	3	17	7	2	14	15		11	4	16	1	4	4	2	8	11	4	12
10	12	4	14	6	17	15	1	8	13		17	7	12	14	1	13	10	2	18	1
11	10	19	6	17	1	3	15	7	5	19		15	9	8	17	7	13	19	2	16
12	17	2	18	1	12	17	3	5	17	11	14		18	18	14	18	17	5	13	12
13	4	9	9	10	14	6	15	11	4	9	7	16		6	4	3	6	11	8	7
14	4	9	5	13	7	6	11	18	8	9	8	14	5		10	3	5	8	8	9
15	8	1	10	2	12	11	6	2	9	3	11	4	7	8		10	10	4	7	4
16	1	8	1	13	11	5	9	18	1	7	4	16	1	2	7		2	11	5	7
17	4	4	4	6	5	8	3	11	10	6	10	9	7	6	10	8		10	10	9
18	9	16	9	18	16	8	8	16	4	18	5	17	16	16	17	18		19	2	
19	11	15	6	13	2	3	18	5	5	17	2	10	10	10	10	9	13	17		16
20	12	4	13	2	18	12	2	15	12	1	14	5	10	12	2	11	13	1	16	

Note.—Tied values have been assigned the highest rank involved (smallest numerical value).

i.e., those searched for in Rank Order Typal Analysis when only pure categories are sought.

Let n equal the number of objects in a submatrix being examined to see whether or not the set of objects qualifies as a type. A type is a submatrix of objects of such a nature that every object in the submatrix has no rank greater than $n - 1$ with any other object of the submatrix (where the ranks are taken from the original matrix).

In order for a pair of objects to qualify under this definition, neither object of the pair has a rank above one with the other object. If the two objects are represented by i and j , i has a rank of one with j , and j has a rank of one with i , and this outcome constitutes a reciprocal pair.

Every matrix contains at least one reciprocal pair. Suppose i and j are the objects between which the highest entry in a matrix mediates. Object i is then highest with j and j is highest with i ; these two objects constitute a reciprocal pair by definition.

Table 2 contains four reciprocal pairs, viz., Objects 3-6, 3-16, 4-12, and 10-20.

Typal triads. In a search for typal triads every object must have no rank higher than 2 (i.e., $n - 1$) in either its column or row. Select any Object i of any reciprocal pair. Let it be Object 3 of Reciprocal Pair 3-6. If Pair 3-6 expands into a type of three objects, it must incorporate the one object second most like Object 3. Object 3 has no object second most like it because two objects, 6 and 16, are tied for being most like it, and both, therefore, have a rank of one with it. Since Object 6 is a member of the reciprocal pair which is being tested for expansion into a triad, this leaves Object 16 to be selected for the test. (If there had been more ranks tied at a value of one, all of their objects would have been tested one at a time with the order being immaterial.) Objects 3, 6, and 16 are assembled in Table 3, using their ranks from Table 2. They do not constitute a type because there is a rank higher than 2 in the matrix, viz., the rank in Row 6 and Column 16. Pair 3-6 cannot be expanded into a type of 3 objects and for this reason it cannot be expanded into a type of more than 3 objects *under the strict definition*. In order for a category of any size to qualify under the strict definition all smaller categories must qualify.

If Pair 3-16 expands into a type of three objects, it must include Object 6. This gives the same non-qualifying triad reported in

TABLE 3
Testing Pairs 3-6 and 3-16 for Expansion

	5 ^b 1 ^a	5 1	5 1	5 3	7	10	10	9	8	13	15	
	3	6	16	9	11	19	1	13	14	17	5	15 ^c
3		1	1	1	3	3	2	5	1	1	2	4
6	1		3	1	1	1	8	3	3	2	5	14
16	1	5		1	4	5	1	1	2	2	11	7
9	3	2	2		4	4	2	1	4	8	7	4
11	6	3	7	5		2	10	9	8	13	1	17
19	6	3	9	5	2		11	10	10	13	2	10
1	8	9	3	5	9	10		4	4	2	4	10
13	9	6	3	4	7	8	4		6	6	14	4
14	5	6	3	8	8	8	4	5		5	7	10
17	4	8	8	10	10	10	4	7	6		5	10
5	10	9	14	11	6	5	7	15	10	10		19
15	10	11	10	9	11	7	8	7	8	10	12	

^a Size of the criterion under which the object qualifies; Objects 3 and 6, for example, qualify as a pair under a criterion of 1 ($n - 1 = 1$).

^b Objects 3, 6, 16, and 9 form a tetrad under a criterion of 5.

^c Assigned to a type of Table 6 under a criterion of 9 or less.

Table 3. Pair 3-16 cannot be expanded under the strict definition into a larger type.

Object 4 of Reciprocal Pair 4-12 has Objects 2, 7, 15, and 20 all tied for second highest with it. They are all tested, one at a time, in Table 4 for expansion of Type 4-12. They introduce one, two, three, and three ranks, respectively, above $n - 1 = 2$. None of them qualifies, and Type 4-12 cannot be expanded under the strict definition into a type of 3 or more objects.

TABLE 4
Testing Pair 4-12 for Expansion

	1 ^a	1	4	6						
	4	12	2	7	15	20	17	10	18	5
4		1	4	6	7	5	13	7	5	10
12	1		2	3	14	12	17	11	5	12
2	2	2		3	2	5	8	5	7	15
7	2	2	2		7	3	6	2	3	15
15	2	4	1	6		4	10	3	4	12
20	2	5	4	2	2		13	1	1	18
17	6	9	4	3	10	9		6	10	5
10	6	7	4	1	1	1	10		2	17
18	9	5	9	8	16	2	18	4		18
5	6	7	16	17	19	18	10	16	17	

^a Size of the criterion under which the object qualifies.

The only other reciprocal pair, 10-20, is tested in Table 5 for expansion into a type of 3 or more objects. It fails to qualify because Object 7, the only object second highest with Object 10, introduces one rank above $n - 1 = 2$. Even though Object 18 is second highest with Object 20 (the other object of Reciprocal Pair 10-20), it need not be tested; its inclusion as one of the 3 objects would exclude Object 7 which is the *only* object second most like Object 10, and Object 18 would have a rank larger than 2 with Object 10. If, on the other hand, Object 18, the only object second highest with Object 20, had been tested first in lieu of Object 7, then Object 7 would not have required a test.

In the classification, thus far, Types 3-6, 3-16, 4-12, and 10-20 have been realized; no other types of the same size can be realized and no larger types can be realized under the strict definition of a type.

Relaxing the criterion. The criterion for typal membership is increased successively from one by steps of one and is applied in expanding the types until all objects of Table 2 are classified into one of two major types or have proven that they cannot be classified. Each successive criterion is applied to every type being built before the criterion is relaxed further. An object satisfies the criterion if it has a rank equal to or smaller than the criterion with an object of the type being tested for expansion.

TABLE 5
Testing Pair 10-20 for Expansion

	1*	1	3	7	5	7							
	10	20	7	15	18	2	4	12	17	16	3	9	13
10		1	1	1	2	4	6	7	10	13	14	13	12
20	1		2	2	1	4	2	5	13	11	13	12	10
7	2	3		7	3	2	2	2	6	11	12	14	15
15	3	4	6		4	1	2	4	10	10	10	9	7
18	4	2	8	16		9	5	18	17	16	16	17	17
2	5	5	3	2	7		2	2	8	14	14	17	13
4	7	5	6	7	5	4		1	13	16	16	15	13
12	11	12	3	14	5	2	1		17	18	18	17	18
17	6	9	3	10	10	4	6	9		8	4	10	7
16	7	7	9	7	11	8	13	16	2		1	1	1
3	11	9	9	4	8	9	13	10	1	1		1	5
9	11	12	14	4	11	17	17	16	8	2	3		1
13	9	7	15	4	11	9	10	16	6	3	9	4	

* Size of the criterion under which the object qualifies.

A comprehensive approach would require a search for new reciprocal pairs each time the criterion is relaxed. When it is relaxed from 1 to 2, a reciprocal pair qualifies if i is first or second most like j and j is first or second most like i . This step was not included in this pencil and paper example because it is too elaborate; it could be included in an electronic computer analysis.

Expansion of the criterion could be stated in a fashion relative to the size of the type being tested for expansion. Under this approach, the criterion would be $n - 1$ initially and then relaxed successively to n , $n + 1$, $n + 2$, etc. The present study uses the amount of absolute discrepancy, as outlined above, on the basis of the assumption that the purpose is to minimize the absolute amount of error in classifying every object.

Under the absolute approach, the criterion for the present data is now increased from 2 to 3 and Object 7 qualifies to extend Type 10-20 into a Triad of 10-20-7, as shown in Table 5. The number three is placed over Object 7 in Table 5 to record that it qualifies under a criterion of three.

Under the criterion of three, Object 10 brings in Object 15 for a test; Object 20 brings in Object 18, and Object 7 brings in Objects 2, 12, and 17, as shown in Table 5. None of them qualifies under a criterion of three.

The criterion of three is applied to Pairs 3-6 and 3-16. Object 3 brings in Object 9 on a test basis; Object 6 brings in Objects 9, 11, and 19, and Object 16 brings in Objects 1, 9, 13, and 14, as shown in Table 3.

Object 9 joins Pair 3-6 to form a triad under the criterion of three; it also joins 3-16 to form another triad under the same criterion. No other types can be formed with either Pairs 3-6 or 3-16 without a larger criterion. Object 9 brings in no other object under this criterion.

Under the criterion of three, neither object of Pair 4-12, Table 4, brings in additional objects (over and above Objects 2, 7, 15, and 20 brought in for tests under a criterion of two), and no additional types are formed from the objects already brought in for tests under a criterion of two.

The criterion is now raised to four. Objects 4 and 12 still bring in no additions. However, Object 2 joins Pair 4-12 to form a triad, 4-12-2, under this criterion, and no other types are formed. Object

2 brings in Object 17 under the criterion of four, and no new types are formed.

Under the criterion of four, Object 3 of triads 3-6-9 and 3-16-9, Table 3, brings in Object 17. No additional types are formed.

Under the criterion of four, none of the objects of Triad 10-20-7, Table 5, brings in any additional objects, and no new types are formed.

The criterion is raised to five. Objects 10 and 7 of triad 10-20-7 bring in no additional objects. Object 20 brings in Object 4. Object 2 joins the triad to form a tetrad of 10-20-7 and 2. Object 2 brings in Objects 4 and 12. No new types are formed.

The criterion is raised to six. Only Object 4 of Triad 4-12-2 brings in an object, viz., 5, and only Type 4-12-2-7 is formed. Object 7 brings in no other objects under a criterion of six.

The criterion of six does not change Tables 3 and 5.

The criterion is raised to seven. Only Object 10 of Tetrad 10-20-7-2, Table 5, brings in an object, viz., 16. Objects 15 and 4 join the type to yield Type 10-20-7-2-15-4. Object 15 brings in Objects 3, 9, and 13, and Object 4, through its associations in Table 4, brings in all of the objects of Table 4.

Table 4 is, therefore, next examined under the criterion of seven. No objects are brought in, and no types are formed.

Tables 4 and 5 are combined in Table 6, listing the members of the two types next to one another.

Under the criterion of seven, Type 3-6-9-16 incorporates Object 11, and Object 11 brings in Object 5, but no additional types are formed.

The criterion is increased to eight. Object 14 of Table 3 qualifies to yield Type 3-6-16-9-11-14. Object 14 brings in Object 15; no additional types are formed.

The criterion of eight brings neither additional objects nor types to Type 10-20-7-2-15-4-12 of Table 6.

The criterion is increased to nine. Objects 10 and 20 each bring in Object 14. Object 2 brings in Object 8. No additional types are formed.

No additional objects are brought into Table 6 by the criterion of nine. Object 13 enters as a typical member in Table 3.

The classification could proceed under repetition of the above steps until the analysis is completed. This would be appropriate with

TABLE 6

Testing Combined Tables 4 and 5 for Further Expansion

	7 ^b															
	1 ^a	1	3	5	7	1	1	16	18		19					
	10	20	7	2	15	4	12	18	17	16 ^a	5	3 ^a	9 ^a	13 ^a	14 ^a	
10		1	1	4	1	6	7	2	10	13	17	14	13	12	14	8
20	1		2	4	2	2	5	1	13	11	18	13	12	10	12	15
7	2	3		2	7	2	2	3	6	11	15	12	14	15	12	10
2	5	5	3		2	2	2	7	8	14	15	14	17	13	15	3
15	3	4	6	1		2	4	4	10	10	12	10	9	7	8	2
4	7	5	6	4	7		1	5	13	16	10	16	15	13	16	3
12	11	12	3	2	14	1		5	17	18	12	18	17	18	18	5
18	4	2	8	9	16	9	5		18	17	18	16	16	17	16	8
17	6	9	3	4	10	6	9	10		8	5	4	10	7	6	11
16	7	7	9	8	7	13	16	11	2		11	1	1	1	2	18
5	16	18	17	16	19	6	7	17	10	14		10	11	15	10	1
3	11	9	9	9	4	13	10	8	1	1	2		1	5	1	11
9	11	12	14	17	4	17	16	11	8	2	7	3		1	4	15
13	9	7	15	9	4	10	16	11	6	3	14	9	4		6	11
14	9	9	11	9	10	13	14	8	5	3	7	5	8	5		18
8	17	19	19	9	17	10	10	16	19	19	7	19	19	19	19	

^a Size of the criterion under which the object qualifies.^b The criterion under which Tables 4 and 5 were combined through Object 4.^c Assigned to a type of Table 3 under a criterion of 9 or less.

an electronic computer and a computer program. However, in a pencil and paper analysis, the steps can be shortened.

All objects have now been assigned to either Table 3 or 6. Objects of Table 3 which have already been assigned by a criterion of nine or less to types of Table 6 are indicated, and analogously, for objects of Table 6 with respect to Table 3.

Table 3 is analyzed observationally to determine the criteria value at which its unassigned objects (unassigned in either Tables 3 or 6) can enter as typical members. Object 1, for example, enters under a criterion of ten; the entrance criterion of each unassigned object is reported over its object code number in Table 3. The same procedure is applied to Table 6. Objects are assigned to the types with which they have their lowest criterion number. In the case of any object not listed in both tables, such as Objects 8 and 18, a check is made with Table 2 in relation to the criterion of the table in which the object is assigned. Object 18 has a criterion assignment of 16 in Table 6. Table 2 shows that it would have a criterion assignment of 19 if it were transferred to Table 3. It remains in Table 6.

Object 8 has a criterion of 19 in Table 6, and it would have a

criterion of the same size in Table 3. Since this is the highest rank possible, Object 8 is in a sense unassignable. It is, however, assigned to Table 6 because it has fewer ranks of nineteen there even when adjusted for the fewer objects in types of that table.

The Hierarchical Classification

The results of the classification are shown graphically in Figure 1. Categories 3-6-9 versus 3-16-9 are optional, and the conflict is resolved at the next higher level where all four objects enter a single category. The chart shows the criterion levels under which objects classify and, thus, gives an indication of the validity with which they enter.

Summary

This paper develops and illustrates a method of classifying fallible data into larger and larger internally consistent categories. Each set of data usually starts with two or more categories which are built up gradually and which combine at various levels. The classification is realized by relaxing gradually and minimally, step by step, the objective criterion of internal consistency.

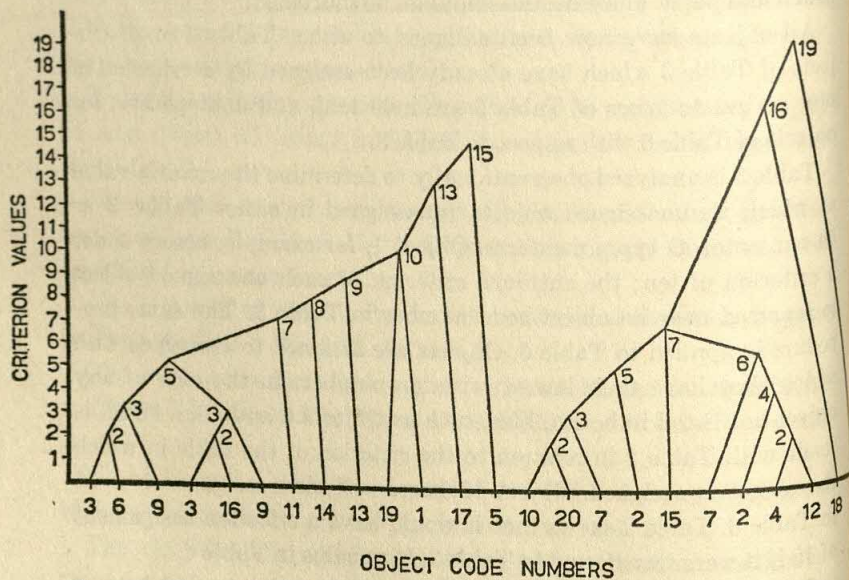


Figure 1. Hierarchical classification by Relaxed Rank Order Typal Analysis.

REFERENCES

- McQuitty, L. L. Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1957, 17, 207-229.
- McQuitty, L. L. Rank order typal analysis. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1963, 23, 55-61.
- McQuitty, L. L. Capabilities and improvements of linkage analysis as a clustering method. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1964, 24, 441-456.
- McQuitty, L. L., Price, L., and Clark, J. A. The problem of ties in a pattern analytic method. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 787-796.

THE STABILITY COEFFICIENT

EDWARD E. CURETON

University of Tennessee

In a previous paper (Cureton, 1958) I gave a formula for the stability coefficient, and quoted the result in a later paper (Cureton, 1965). I did not know then that essentially the same formula had been given previously by Remmers and Whistler (1938). The formula is correct, but both my derivation and the one given by Remmers and Whistler were slightly defective. A derivation that seems to me to be more nearly correct is given below, together with some further discussion.

The *inconsistency* of a test may be defined as that part of its error of measurement which is associated with items or forms. One set of items can never draw forth from an examinee a set of reactions completely representative of the totality of potential reactions which represent the ability or trait measured by the universe of items of which this set (test form) is a random or stratified-random sample. Hence another set (form), similarly selected from the same item universe but differing in specific content, will in general yield a somewhat different score. Most of the literature on test reliability is concerned exclusively or almost exclusively with consistency. Two forms of a test are *parallel* if their items are parallel random or stratified-random samples from the same item universe.

The *instability* of a test is that part of its error of measurement which is associated with the particular time and occasion on which it is administered. The ability or trait measured by a test, quite apart from learning, mental growth, permanent forgetting, or mental decline, fluctuates with time. These fluctuations include reactions

to session differences in the examiner's procedures, the working conditions at each particular session, and random and cyclic variations in emotional control, general fatigue, motivation, attitude toward the test, anxiety, working procedures including working speed, resistance to distraction, and access to memory, to name but a few. The true score of an examinee would then be defined as his average score both on different forms of the test and on different examining occasions. As regards these occasions, however, we would like to assume that over the time span involved there is no mental growth, no mental decline, no learning, and no forgetting: this last defined as loss of retention. Fluctuations in ability to recall, given constant retention, are elements of instability.

For a long test, instability might be appreciable over a time as short as the time required to complete it, so that serial administration of two forms at the same sitting is not necessarily equivalent to simultaneous administration (as by, say, the odd and even items of one double-length form). On the other hand, cyclic variations in some of the elements of instability come in very long cycles. Differential seasonal variation (differential over persons) is at least a tenable hypothesis, and euphoric-depressive cycles within the normal range may be even longer. It would seem, then, that no practical interval between the administration of two forms of a test is long enough to provide assurance that the two sets of instability errors are uncorrelated.

At least equally important is the fact that true-score change is a continuous process also. Learning occurs as an examinee proceeds from the first to the last item of a single test, and the amount learned is different for different examinees. Between test sessions, each examinee is forgetting whenever he is not learning, and whenever he is learning in areas which are irrelevant to the area tested. If, in addition, the interval between test sessions is substantial, differential mental growth may be appreciable in children, and differential mental decline in older adults. Hence no interval is short enough to rule out the possibility that differential true-score changes have occurred. So far as measurement is concerned, fluctuations in the function measured are *always* contaminated by changes in the true level of the function. Instability, as best we can measure it, is an unknown combination of function fluctuation and function change.

Theory

Using the linear model of weak true-score theory, we can write,

$$x_1 = k_1(x_a + s_1) + e_1, \quad (1)$$

$$x_2 = k_2(x_b + s_2) + e_2.$$

Here x_1 and x_2 are raw scores on the two forms of the test, x_a and x_b are the corresponding true scores on the occasions when the two forms are administered, s_1 and s_2 are the instability errors on the two occasions, e_1 and e_2 are the inconsistency errors of the two forms, and k_1 and k_2 are constants associated with possible inequalities in the raw-score units of measurement, and hence in the reliabilities and variances of the two forms. We can assume without loss of generality that all scores are taken as deviations from their means as origins, that x_a , x_b , s_1 , and s_2 are measured on the same true-score scale, and that e_1 and e_2 are measured on the raw-score scales of x_1 and x_2 respectively. We assume that x_a and x_b are uncorrelated with s_1 and s_2 , that all four of these are uncorrelated with e_1 and e_2 , and that e_1 and e_2 are uncorrelated with each other. We do not assume that x_a is uncorrelated with x_b , or that s_1 is uncorrelated with s_2 .

With this model,

$$r_{12} = k_1 k_2 (\sigma_{ab} + \sigma_{s_1 s_2}) / \sigma_1 \sigma_2. \quad (2)$$

Here σ_{ab} is the true-score covariance, and $\sigma_{s_1 s_2}$ is the covariance of the instability errors. By the usual variance ratio definition, the consistency coefficients are

$$C_1 = k_1^2 (\sigma_a^2 + \sigma_{s_1}^2) / \sigma_1^2, \quad (3)$$

$$C_2 = k_2^2 (\sigma_b^2 + \sigma_{s_2}^2) / \sigma_2^2.$$

We will assume that over the interval between the administration of the two forms, neither the true-score variance nor the variance of the instability errors changes appreciably; i.e., that

$$\sigma_a^2 = \sigma_b^2 = \sigma_x^2,$$

$$\sigma_{s_1}^2 = \sigma_{s_2}^2 = \sigma_s^2.$$

Then (3) becomes

$$C_1 = k_1^2 (\sigma_x^2 + \sigma_s^2) / \sigma_1^2, \quad (4)$$

$$C_2 = k_2^2 (\sigma_x^2 + \sigma_s^2) / \sigma_2^2,$$

and

$$\sqrt{C_1 C_2} = k_1 k_2 (\sigma_x^2 + \sigma_{e_1}^2) / \sigma_1 \sigma_2. \quad (5)$$

Then dividing (2) by (5),

$$r_s = \frac{r_{12}}{\sqrt{C_1 C_2}} = \frac{\sigma_{ab} + \sigma_{e_1 e_2}}{\sigma_x^2 + \sigma_{e_1}^2}. \quad (6)$$

From the second expression of (6) it is clear that r_s is of the form of a correlation corrected for attenuation: it is the interform correlation corrected for the attenuation due to the inconsistencies of the two forms. It seems reasonable to term r_s the *stability coefficient* of the test over the interval between the administration of the two forms.

For computation, r_{12} is the product-moment correlation between the two forms, and C_1 and C_2 may be computed by the Kuder-Richardson Formula 20 or by the split-half method and the Spearman-Brown formula.

If there were no changes in true scores over the interval between the administration of forms 1 and 2 of the test, we could say

$$x_a = x_b = x,$$

and (6) would become

$$r_s = \frac{r_{12}}{\sqrt{C_1 C_2}} = \frac{\sigma_x^2 + \sigma_{e_1 e_2}}{\sigma_x^2 + \sigma_{e_1}^2}. \quad (7)$$

But since the second expression of (7) is the same as that of (6), it is clear that we have no way to determine empirically whether σ_{ab} is appreciably lower than σ_x^2 or not. As the length of the interval increases, $\sigma_{e_1 e_2}$ will decline from σ_x^2 , its value for simultaneous administration, to zero when the instability errors on the second occasion are random with respect to their values on the first occasion; and thereafter it will probably fluctuate between zero and some low positive value associated with the time of day, the day of the week, the season of the year, etc., of the second occasion, as compared to the first occasion. The value of σ_{ab} will decline continuously from σ_x^2 , its value for simultaneous administration, but for most tests it should never reach zero. Whether these declines are most rapid at first and then flatten out, or are inverted S-curves, can only be determined empirically.

From the preceding discussion, it is apparent that in reporting a stability coefficient, the time interval between the first and second

administrations should be noted explicitly, and ideally the time of day and day of the week of each administration, and any intervening events (e.g., examinations, snowstorms, epidemics, athletic events, and the like) which might affect differential instability to a greater than average degree. If the time interval is long, any known relevant learning experiences, affecting some examinees but not others, which occur between the first administration and the second should be noted also. The stability coefficient is always attenuated in greater or lesser degree by true-score changes, and is a function of the *specific* interval between the first administration and the second, not merely of the length of this interval.

Since the consistency coefficients enter into the formula for the stability coefficient, both forms of the test should be administered ideally without time limits, and in practice with time limits sufficient to permit at least 90 per cent of the examinees to attempt every item. In the latter case every examinee's answers should be augmented by a random answer for every item omitted, before scoring.

Note that in (6) and (7) the values of k_1 , k_2 , σ_1 , and σ_2 do not enter; they cancel when we divide (2) by (5). From this it follows that while forms 1 and 2 must be parallel forms in the sense that they both measure the same function or combination of functions, they need *not* be *equivalent*: they do not have to be equally reliable or equally variable, or to measure in comparable raw-score units. One form may be a long form and the other a short form, so long as the two forms consist entirely of different items measuring the same function.

When the split-half method is used to compute the consistency coefficients, r_s may be computed by formulas alternative to (6). The model is

Half-Test	Occasion and Test Form	
	I	II
A	x_1	x_2
B	x_3	x_4

Then

$$r_s = \frac{r_{12} + r_{14} + r_{23} + r_{34}}{4\sqrt{r_{13}r_{24}}}; \quad (8)$$

$$r_s = \frac{\sqrt{r_{12}r_{14}r_{23}r_{34}}}{\sqrt{r_{13}r_{24}}}. \quad (9)$$

In general, (9) is the preferred formula, since it does not require that either x_1 and x_3 or x_2 and x_4 be equally reliable or equally variable. But if these conditions do hold, (8) appears to be at least as good as (9) and perhaps slightly better. If (6) is used, with C_1 and C_2 computed by the split-half method and the Spearman-Brown formula, the two split-halves of each test form must be equally reliable and equally variable, and (6) is algebraically equivalent to (8).

The test-retest coefficient has often been assumed to be a stability coefficient because the test items are the same on both occasions. If we start with (1), we can reasonably set $k_1 = k_2$, and hence omit the k 's from these equations: the same form should measure in the same units (and hence be equally reliable and equally variable) on two different occasions. We then have

$$x_1 = x_a + s_1 + e_1, \quad (10)$$

$$x_2 = x_b + s_2 + e_2,$$

$$r_{12} = (\sigma_{ab} + \sigma_{s_1s_2} + \sigma_{e_1e_2})/\sigma_1\sigma_2. \quad (11)$$

The argument based on identical items would seem to imply that $r_{e_1e_2} = 1$, and hence that $\sigma_{e_1e_2} = \sigma_e^2$. Then assuming again that $\sigma_a^2 = \sigma_b^2 = \sigma_s^2$ and $\sigma_{s_1s_2} = \sigma_{s_2s_2} = \sigma_s^2$, the variances are

$$\sigma_1^2 = \sigma_2^2 = \sigma_s^2 + \sigma_e^2 + \sigma_e^2,$$

and

$$r_{12} = \frac{\sigma_{ab} + \sigma_{s_1s_2} + \sigma_e^2}{\sigma_s^2 + \sigma_e^2 + \sigma_e^2}. \quad (12)$$

It is hard to see how (12) can be interpreted as a stability coefficient: σ_e^2 , the inconsistency-error variance, appears in both the numerator and the denominator, and does not cancel out. If we compare (12) with (6), it is evident that even under the assumption that $r_{e_1e_2} = 1$, (12) will give an inflated estimate of the stability, and the amount of this inflation will vary *inversely* with the *consistency* of the test.

The assumption that $r_{e_1e_2} = 1$ is fallacious on two counts. First, there will be perseveration and perhaps proactive inhibition effects from the first test session to the second. Memory at the second session

of particular item responses given at the first session will be only a part of these effects. But even apart from these considerations, we cannot assume that $e_1 = e_2$ for every examinee. For many items, the probability that an examinee will mark the right answer is neither 0 nor 1. Then quite apart from any over-all differences in test-taking ability represented by s_1 and s_2 , an examinee may mark different answers to the same item on two occasions even when the probabilities remain constant. Hence, on both counts, (12) must be replaced by

$$r_{12} = \frac{\sigma_{ab} + \sigma_{s_1, s_2} + \sigma_{e_1, e_2}}{\sigma_s^2 + \sigma_{s_1}^2 + \sigma_{s_2}^2}, \quad (13)$$

and (13) may give either an inflated or an attenuated estimate of the value given by (6), depending upon the unknown magnitudes of the perseveration and item-probability effects.

If we divide (11) by (4), with the k 's omitted from the latter, we obtain

$$r_{12} = \frac{\sigma_{ab} + \sigma_{s_1, s_2} + \sigma_{e_1, e_2}}{\sigma_s^2 + \sigma_{s_1}^2}. \quad (14)$$

While $\sigma_{e_1, e_2} < \sigma_s^2$, it is almost certain that it will remain positive, in which case (14) will always give an inflated estimate of the stability coefficient. So far as I am aware, no one has proposed (14) as a formula for the stability coefficient, and I certainly do not propose it here.

So far as I can tell, the test-retest coefficient has no clear interpretation under weak true-score theory. It depends upon a conglomeration of instability errors, inconsistency errors, perseveration effects, and item-probability effects, whose relative contributions to its magnitude cannot be untangled. The only exception would seem to be a pure speed test (such as, e.g., a number checking test), with the second administration coming some weeks or months after the first. In such a case, it would perhaps be not unreasonable to interpret the test-retest coefficient as an inter-form reliability coefficient.

Data

We consider first the proposition that instability may be appreciable over a period as short as that required by a group of examinees to take one test.

A final examination of 60 four-choice items was given to a large class in elementary psychology. To minimize copying, students

sitting in alternate seats took a "green" form (Form G) and a "white" form (Form W). The two forms consisted of the same questions arranged in different orders. In both forms the items were arranged essentially randomly with respect to difficulty, discrimination, and topic. The analyses of the two forms may be considered replications, with the same items but with different subjects and different split-halves.

For each group, four scores were obtained, each on 15 items:

- x_1 : odds of first 30
- x_3 : evens of first 30
- x_2 : odds of second 30
- x_4 : evens of second 30

Thus r_{13} and r_{24} will be consistency coefficients, while r_{12} , r_{14} , r_{23} , and r_{34} will be inter-form correlations. The stability coefficients were computed by (9). The results are

	White form	Green form
N	726	734
$C = \sqrt{r_{13}r_{24}}$.5178	.4825
$r = \sqrt{r_{12}r_{14}r_{23}r_{34}}$.5040	.4918
$r_s = r/C$.973	1.019

The proposition is not proved by these data, even with samples of over 700. One possible reason is that without stratification by topic, the subforms were somewhat lacking in parallelism.

A better test, for stability over a one to three day period, is provided by data reported by Davis (1968). Two carefully constructed forms of a test measuring eight aspects of reading comprehension, each with 12 items, were given at intervals ranging from one to three days to 988 twelfth-grade students. Consistency coefficients were computed from the odd and even items of each form by a modification of a formula due to Angoff (1953), which reduces to the Spearman-Brown formula if the odds and evens have equal variance. From his results we have, using (6) to compute values of r_s .

With the one exception of Test 2, the stability coefficients are clustered in the region between .89 and .96, and the difference between 1.067 and the next highest (.110) is almost twice as great

Test	r_{11}	r_{22}	r_{12}	$\sqrt{r_{11}r_{22}}$	r_s
1	.689	.602	.582	.6440	.904
2	.557	.644	.639	.5989	1.067
3	.644	.713	.621	.6776	.916
4	.632	.717	.644	.6732	.957
5	.660	.671	.594	.6655	.892
6	.656	.670	.633	.6630	.955
7	.728	.676	.665	.7015	.948
8	.754	.677	.677	.7145	.948
Mean	.665	.671	.632	.666	.948

as the distance between the next highest and the lowest (.065). If we regard the 1.067 as anomalous, the mean of the other seven is .931.

While tests of the hypothesis that a single correlation corrected for attenuation is not significantly different from unity exist (Forsyth and Feldt, 1969; Lord, 1957; McNemar, 1958), there is no test for the hypothesis that the mean of seven or eight such coefficients, all correlated, does not differ significantly from unity. In view of the sample size, however (988), it seems safe to conclude that some instability exists over the one to three day period.

In a previously reported study, (Cureton, 1939, 1965), three 27-item forms of the A.C.E. Opposites Test were mimeographed on one sheet and administered consecutively to 187 undergraduate students in five classes on a Friday. Three other forms, mimeographed also on one sheet, were given to some of the same classes on the following Monday, to some on the following Wednesday, and to some on the following Friday. All were given without time limit. The average of the six within-sheet correlations was .729; the average of the nine across-sheet correlations was .500.¹ If we regard the within-sheet correlations as estimates of consistency, the stability coefficient is $.500/.729 = .686$.

One other study, also previously reported (Cureton, 1965; Loveland, 1952), included the Verbal Reasoning Test of the D.A.T., Form A. This test, along with all others of the D.A.T. except

¹Incorrectly reported as .517 in Cureton (1965).

Clerical Speed and Accuracy, was administered to 572 students in grades 9 through 12 of one high school, and the *same form* was re-administered eight days later. Odd and even items were scored separately for each administration. We have the model,

Form	Occasion	
	I	II
Odd	x_1	x_2
Even	x_I	x_{II}

Here r_{1I} and r_{2II} are consistency coefficients, r_{1II} and r_{I2} are inter-form correlations contaminated by differential practice effects, and r_{12} and r_{II} are test-retest coefficients. The data are

$$C = \sqrt{r_{1I}r_{2II}} = .849$$

$$r = (r_{1II} + r_{I2})/2 = .805$$

$$r_s = r/C = .949$$

$$r_{t-r} = (r_{12} + r_{II})/2 = .846$$

The stability coefficient is appreciably higher than the value for the A.C.E. Opposites Test, and is in fact higher than the mean value for the Davis reading tests when the anomalous 1.067 is omitted, even though the interval was eight days as against averages of about five days and two days. How much of this may be attributed to contamination due to differential practice effects and use of the same form on both occasions is not clear. Note also that r_{t-r} , the test-retest coefficient, is much lower than r_s . If we regard it as an interform correlation, however, and divide it by C , the result is .996, which is almost surely a gross overestimate of the stability.

These four studies deal with different types of material, but all of them, at least, are verbal tests. We can estimate the average intervals for the four studies as about one hour, two days, five days, and eight days. For consecutive testing the instability is near zero, for the two-day interval it is slight, for the five-day interval it is substantial, and for the eight-day interval, due in some part to contamination, it is again slight. We can conclude *very tentatively* that for verbal tests the instability curve is of the inverted-S type, with instability increasing slowly at first, then faster, and

finally more slowly again. All four of these studies probably represent only a part of the interval over which instability is increasing at an accelerated rate.

REFERENCES

- Angoff, W. H. Test reliability and effective test length. *Psychometrika*, 1953, 18, 1-14.
- Cureton, E. E. Note on the validity of the American Council on Education Psychological Examination. *Journal of Applied Psychology*, 1939, 23, 306-307.
- Cureton, E. E. The definition and estimation of test reliability. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1958, 18, 715-738.
- Cureton, E. E. Reliability and validity: Basic assumptions and experimental designs. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1965, 25, 327-346.
- Davis, F. B. Research in comprehension in reading. *Reading Research Quarterly*, 1968, 3, 499-545.
- Forsyth, P. A. and Feldt, L. S. An investigation of empirical sampling distributions of correlation coefficients corrected for attenuation. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 61-67.
- Lord, F. M. A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. *Psychometrika*, 1957, 22, 207-220.
- Loveland, E. H. *Measurement of factors affecting test-retest reliability*. Unpublished Ph.D. thesis, University of Tennessee, 1952.
- McNemar, Q. Attenuation and interaction. *Psychometrika*, 1958, 23, 259-266.
- Remmers, H. H. and Whistler, L. Test reliability as a function of method of computation. *Journal of Educational Psychology*, 1938, 29, 81-92.

INTEGRATION OF CONCEPTS OF RELIABILITY AND STANDARD ERROR OF MEASUREMENT

JOHN L. HORN¹
University of Denver

WHAT are the assumptions underlying derivations of various indices of error of measurement and such coefficients of reliability as KR-20 and KR-21? Over the years there has been a great variety of instructive discussion relating to questions of this kind (cf. Cronbach, 1951; Cronbach, Rajaratman and Gleser, 1963; Gulliksen, 1950; Henrysson, 1959; Hoyt, 1941; Kuder and Richardson, 1937; Lord 1955a; 1955b; 1957; 1959a; 1959b; 1962; Novick, 1966; Novick and Lewis, 1967; Penfield, 1967; Winer, 1962 and others). Recent articles by Lord have been particularly helpful in indicating the minimal assumptions under which a reliability coefficient may be obtained and in showing the basis upon which one would need to justify use of one standard error of measurement (*SEM*) for all scores. Yet it is true that interesting and important relationships between various derivations still remain somewhat obscure and the implications in use of various ways of estimating a standard error of measurement are by no means evident. The purpose of this paper is to explicate some of the problems implied by these statements and to indicate some of the practical implications of various proposed solutions. This may also point the way toward a clearer conceptualization of basic theoretical issues in this area.

Random Response Models

One of the more logically compelling ways of defining an *SEM* and a reliability coefficient (r_{tt}) is in terms of a classical (Fisher-

¹I wish to thank Professors Bernard Spilka and Lloyd G. Humphreys and Mrs. Betty Rossman for suggestions which are contained in several of the ideas developed in this paper.

ian) inferential model. Perhaps the first clear indication of this kind of definition was given by Hoyt (1941), but Lord (1955a; 1955b; 1957), Penfield (1967) and Winer (1962) have also developed the basic idea. It would appear, however, that we should consider two somewhat different variations on this theme. These correspond to the distinction between the alpha or KR-20 formula for reliability and the formula referred to as KR-21. But the implications may run deeper than this.

In the most direct approach to reliability via an inferential model, the m items (stimuli) of a test (any measuring procedure) are assumed to be a representative sample drawn from a population of stimuli to which a subject could respond to indicate a magnitude of the attribute in question. It is assumed (for the purpose of rejecting the assumption) that the responses to stimuli could be random. That is, the hypothesis which is to be rejected stipulates that the m variables generated by N subjects responding to m items are random variables. It need not be assumed that all items have exactly the same difficulty level (eccentricity for non-ability items), but the model implies that the variation in eccentricity, as in other statistics for the distribution, will be only that expected by chance. It is assumed that the scale value for a subject will be obtained by simple linear combination of the response scores for the individual items. It presents no theoretical problem to divide scale values by a constant and it is convenient to do this, making the constant m , the number of items. A scale value can then be seen as a mean over m variables drawn representatively from a population of such variables.

If an individual would respond to all stimuli in the population, his true score, τ_i , would be the mean in this population. It is assumed that the mean, t_i , of responses to a sample of stimuli is an unbiased estimate of τ_i . In the random response model there is an assumption that responses to items are drawn independently—selection of one does not determine selection of another. Obviously, however, this is a model assumption which an investigator would hope to be able to reject. That is, he would expect that correlation between items would be significantly larger than zero. If stimuli are drawn from a population of stimuli responses to all of which are indicative of the attribute to be measured, then the correlation between such items will be positive and the random response model

will not adequately represent the behavior in question. An alternative hypothesis can thus gain credence.

Several consequences follow from this conceptualization. First, the variance of item responses around an obtained score, t_i , is seen to be an efficient estimate of the sampling error (in the population of stimuli) for measurements of value, t_i . In other words,

$$\sigma_i = \sqrt{\sum_j \frac{(X_{ij} - t_i)^2}{m - 1}} \quad (1)$$

(where X_{ij} represents the response of an individual i to stimulus j) estimates a standard error of measurement for subjects obtaining the score t_i . This implies that the *SEM* for different measurement values will be (can be and, in general, is) different, as Lord (1957) has emphasized. This idea also has empirical support. For example, there are McNemar's (1942) well-known results showing that high scores on the Stanford-Binet test have a larger error variability than do the low scores.

Second, this statement of the problem leads directly to a test of an hypothesis that measurement has been achieved. If measurement were not achieved, then the implication is that response to one stimulus in the set of m stimuli does not indicate presence of the same attribute as is indicated by response to other stimuli in the set. In other words, this implies that responses are indeed random. If responses to the separate stimuli are random, then the above-mentioned σ_i^2 for each subject is an estimate, based upon a sample of size m , of the variance of a population of such random responses. If it is assumed that sampling of subjects is independent—the selection of one subject does not determine the selection of another—and that the σ_i^2 estimates are homogeneous, then the pooled variance

$$\sigma_{\text{pop-w}}^2 = \frac{\sum_i \sigma_i^2}{N} = \frac{\sum_i \sum_j (X_{ij} - t_i)^2}{N(m - 1)} \quad (2)$$

can be regarded as an efficient estimate of the population variance for the population. This is a variance estimate derived from variability *within* subjects. The subscript *pop-w* in (2) represents this fact.

Another estimate of the population variance can be obtained

from the marginal variance among t_i values. For since any t_i is a mean for a sample of size m in the population,

$$\sigma_i^2 = \frac{\sum_i (t_i - \bar{t})^2}{N - 1} \quad (3)$$

is an estimate of the variance of means; and since the variance of means and the variance of elements have the relationship,

$$\sigma_{\text{means}}^2 = \frac{\sigma_{\text{pop}}^2}{\text{sample size}} \quad (4)$$

(as in the usual developments of analysis of variance) it follows that

$$\sigma_{\text{pop-b}}^2 = m\sigma_i^2 = \frac{m \sum_i (t_i - \bar{t})^2}{N - 1} \quad (5)$$

is an estimate of the population variance.

The estimates of (2) and (5) are independent. The ratio of these independent estimates of a variance

$$F = \frac{\sigma_{\text{pop-b}}^2}{\sigma_{\text{pop-w}}^2} = \frac{m \sum_i (t_i - \bar{t})^2 / (N - 1)}{\sum_j \sum_k (X_{jk} - \bar{t})^2 / N(m - 1)} \quad (6)$$

has an F distribution with $(N - 1)$ and $N(m - 1)$ degrees of freedom.

These developments thus provide a basis for testing an hypothesis counter to the hypothesis of random variation. If the F is significant, the pertinent alternative hypothesis becomes tenable. This stipulates that subjects responses to the various stimuli are *not* independent; that, in fact, responses to one stimulus indicates presence of the same attribute as is indicated by responses to other stimuli. This is a basis, then, for rejecting a claim that the responses do not indicate measurement.

When $\sigma_{\text{pop-b}}^2$ is appreciably larger than $\sigma_{\text{pop-w}}^2$ the hypothesis of equality of these two variance estimates can be rejected. This will occur when the covariance term, $\sum \sum r_{jk} \sigma_j \sigma_k$, in the general expression for the variance of a linear composite variable

$$\sigma_i^2 = \sum \sigma_i^2 + \sum \sum r_{ik} \sigma_i \sigma_k \quad (7)$$

is appreciably larger than zero (cf. Horn, 1963). If the average correlation between responses to stimuli j and k is appreciably larger than zero, then the average of the covariances will be nonzero. The $\sigma_{\text{pop-b}}^2$ will then tend to be large, implying that F will be large and

thus making it likely that a "no measurement" hypothesis can be rejected and an alternative hypothesis—implying measurement—can be accepted. It is on this basis, then, that Lord's development of SEM and reliability theory can mesh with hypothesis testing.

Also, this development clearly indicates that when the sum of the covariances tends to zero, the variance of a composite tends to equal the sum of the variances of the components. Under these conditions if items are dichotomous, with about one-half of the subjects responding in each way to each item, the distribution of the composite scores will be a symmetric binomial that approximates a normal distribution. On the other hand, as the sum of the covariances increases, the variance increases and the distribution form becomes more platykurtic, approaching U-shape as item intercorrelations approach 1.0. Clearly, then, an approximately normal distribution can obtain under conditions of no measurement, whereas departures from normality indicate that one can retain an hypothesis that measurement has been achieved.

Under the assumption of no measurement, the median of the F statistic will be 1.0 or less. It is reasonable to argue that a descriptive statistic which would indicate the extent to which the "no-measurement" assumption does not fit with the data is one in which an obtained value of F is compared with an expected value, as by subtracting the median value from the obtained value— $F-1$. If this is large it indicates a departure from the "no-measurement" conditions of the random-response model. However, the value of $F-1$ is partly a function of the number of degrees of freedom and this would vary from one application to another. To get a statistic that would not vary in this manner one could divide out the degrees of freedom factor, thus

$$r_{tt} = \frac{F - 1}{F} \quad (8)$$

As demonstrated by Lord (1957) if an assumption of dichotomous items is made, the reliability coefficient defined in this manner is closely similar to the coefficient widely known as KR-21.

We can see this by substituting the expression for F into (8), whence we obtain

$$r_{tt} = 1 - \frac{\sigma_{\text{pop-w}}^2}{\sigma_{\text{pop-b}}^2} \quad (9)$$

and then simplifying the within-variance term. To do this note that

$$\begin{aligned} \frac{\sum_i (x_{ij} - t_i)^2}{m-1} &= \frac{\sum x_{ij}^2 - mt_i^2}{m-1} = \frac{\sum x_{ij}^2 - m\left(\frac{\sum x_{ij}}{m}\right)^2}{m-1} \\ &= \frac{m \sum x_{ij}^2 - (\sum x_{ij})^2}{m(m-1)} \quad (10) \end{aligned}$$

where under the assumption that x_{ij} is a dichotomous variable, it can be assumed with no loss of generality that $\sum x_{ij}^2 = \sum x_{ij}$. Also $\sum x_{ij}$ is simply the sum of the item responses—i.e., the total score, which, for convenience, we may represent as T_i . Thus, the variance-within can be seen to reduce to

$$\begin{aligned} \sigma_{\text{pop-w}}^2 &= \frac{\sum_i \left[m \sum_j x_{ij} - (\sum_j x_{ij})^2 \right]}{Nm(m-1)} = \frac{\sum_i (mT_i - T_i^2)}{Nm(m-1)} = \\ &= \frac{m\bar{T} - S_T^2 - \bar{T}^2}{m(m-1)} = \frac{\bar{T}(m - \bar{T}) - S_T^2}{m(m-1)} \quad (11) \end{aligned}$$

(where it is recognized that since $S_T^2 = \sum T_i^2/N - \bar{T}^2$, it follows that $\sum T_i^2/N = S_T^2 + \bar{T}^2$).

The variance-between can be put into a somewhat similar form by observing that the variance for the proportions in (5) is

$$\begin{aligned} \sigma_{\text{pop-b}}^2 &= \frac{m \sum \left(\frac{T_i}{m} - \frac{\bar{T}}{m} \right)^2}{N-1} \\ &= \frac{1}{m} \frac{\sum (T_i - \bar{T})^2}{N-1} = \frac{\sigma_T^2}{m} \quad (12) \end{aligned}$$

Making these substitutions for $\sigma_{\text{pop-b}}^2$ and $\sigma_{\text{pop-w}}^2$ into (9) and assuming that $\sigma_T^2 \cong S_T^2$, we obtain

$$\begin{aligned} r_{ii} &= 1 - \frac{\bar{T}(m - \bar{T}) - S_T^2}{(m-1)S_T^2} \\ &= \frac{mS_T^2 - \bar{T}(m - \bar{T})}{(m-1)S_T^2} = \frac{m}{m-1} \left[1 - \frac{\bar{T}(m - \bar{T})}{mS_T^2} \right] \quad (13) \end{aligned}$$

perhaps the most commonly seen expressions for formula KR-21.

In arriving at this computable representation of a concept of reliability it was assumed that items were dichotomous. However, this assumption was not contained in equation (9), from which the

KR-21 formula was derived, and this, also, was computable. Hence, it can be argued that the basic assumptions involved in this definition of reliability are essentially only those of linear analysis of variance. In the structural model underlying this model there was an assumption that an obtained response was comprised of two components—the true and the error components—which combined additively. In the statement that the variance estimates are independent there was contained an assertion that the expected covariance for the error and true components was zero. The variance between persons could then be seen to contain the variance due to the true score plus error variance, while the pooled within-person variance could be seen to be comprised only of this latter. However, it was not necessary to assume that the variances for items were equal, as sometimes seems to be supposed in discussions of the rationale for the KR-21 formula. Rather, as noted, the assumption of this kind was merely one of homogeneity of item variances. In other words, the KR-21 formula need not involve an assumption that item difficulties are precisely equal; although it can seem to involve this assumption when it is derived algebraically from the KR-20 formula (see Horst, 1966 for a recent example of this derivation).

Hoyt's (1941) well known statement of a concept of reliability derives from a somewhat different set of assumptions than those stated above. In this statement of the problem Hoyt reversed the roles, so to speak, in what is commonly referred to as a single-factor-to-repeated-measures analysis of variance design. That is, instead of regarding the items of a test as treatments, whence it becomes reasonable to use the variance of the residual (σ^2_{res})—what is sometimes referred to as interaction—as the variance-estimate term (error) in testing a main effect due to items, Hoyt treated the subjects as treatments and then went on to use σ^2_{res} as the variance-estimate term in a test for a main effect due to subjects. This leads to a statement of reliability which is formally very similar to that of equation (9), namely

$$r_{tt} = \frac{F - 1}{F} = 1 - \frac{\sigma^2_{res}}{\sigma^2_{pop-b}} \quad (14)$$

but which, in fact, is different in some interesting, if not practically important, ways.

In particular this statement of reliability has the effect of removing variance resulting from differences in item difficulties. At a practical level this means that the reliability calculated by this procedure usually will be larger than the reliability calculated by the procedure of equation (9).

This can be seen by recalling that equation (9) involves a within-person sum of squares which can be partitioned into a term representing the sum of squares of the residual plus a term, usually not zero, representing the between-item variability in eccentricity (difficulty). That is, as is shown in detail by Sheffé (1960)

$$\begin{aligned} \sum_i^N \sum_j^m (x_{ij} - t_i)^2 \\ = \sum_i^N \sum_j^m (x_{ij} - t_i - M_j + \bar{t})^2 + N \sum_j^m (M_j - \bar{t})^2 \end{aligned} \quad (15)$$

$$SS(\text{Within subjects}) = SS(\text{Residual}) + SS(\text{Between Items})$$

$$SS_w = SS_{res} + SSM$$

where M_j represents the mean (over N) for item (treatment) j , SS stands for sum of squares, and the meanings of the other symbols are as defined previously. The $N(m - 1)$ degrees of freedom associated with the within-subjects sum of squares is similarly partitioned into $(m - 1)$ for the between-items term and $(m - 1)(N - 1)$ for the residual. Thus, solving (15) for SS_{res} , the variance estimate based upon the residual can be seen to be

$$\sigma_{res}^2 = \frac{SS_w - SSM}{(N - 1)(m - 1)} \quad (16)$$

In the symbols of this equation σ_{pop-w}^2 of equation (9) is

$$\sigma_{pop-w}^2 = \frac{SS_w}{N(m - 1)} \quad (17)$$

For purposes of comparison it is convenient to multiply on both sides of (16) by $(N - 1)$ and on both sides of (17) by N and then solve for σ_{res}^2 thus

$$\begin{aligned} (N - 1)\sigma_{res}^2 &= \frac{SS_w - SSM}{m - 1} \\ &= N\sigma_{pop-w}^2 - \frac{SSM}{m - 1} \end{aligned} \quad (18)$$

$$\sigma^2_{res} = \frac{N}{(N-1)} \sigma^2_{pop-w} - \frac{SSM}{(N-1)(m-1)}$$

This indicates clearly that it is possible—with small N , thus a large $N/(N-1)$ ratio, and small sum of squares due to differences in item difficulties—for the variance of the residual to be larger than the within-person variance. In fact, if item difficulties were precisely equal (implying that SSM is zero), then σ^2_{res} must be larger than σ^2_{pop-w} by the ratio $N/(N-1)$. More often in the practice of research N will be large, the ratio of N to $(N-1)$ will be very nearly 1.0, some variance in item difficulties will obtain, thus σ^2_{res} will be smaller than σ^2_{pop-w} and the reliability computed by (14) will be larger than the reliability computed by equation (9).

After defining reliability in a manner equivalent to equation (14) Hoyt, observed that: "It may be interesting to some who are familiar with the work of Kuder and Richardson that the foregoing method of estimating the coefficient of reliability gives precisely the same results as formula (20) of their paper. This fact can be easily verified algebraically."

The algebra to which Hoyt refers, although a bit more tedious than one might hope, can be indicated by first recalling that the variance of the residual reduces to

$$\sigma^2_{res} = \frac{\sum \sigma_i^2}{m} - \frac{\sum \sum r_{ik} \sigma_i \sigma_k}{m(m-1)} \quad (19)$$

(Winer, 1962; pp. 120-122). Then notice from equations (15), (7) and (12) that the variance between subjects can be expressed as

$$\sigma^2_{pop-b} = \frac{\sigma_T^2}{m} = \frac{\sum \sigma_i^2 + \sum \sum r_{ik} \sigma_i \sigma_k}{m} \quad (20)$$

With these observations the reliability of equation (14) can be rather directly reduced to

$$\begin{aligned} r_{tt} &= 1 - \frac{\left(\sum \sigma_i - \frac{\sum \sum r_{ik} \sigma_i \sigma_k}{m-1} \right) \frac{1}{m}}{\left(\sum \sigma_i^2 + \sum \sum r_{ik} \sigma_i \sigma_k \right) \frac{1}{m}} \\ &= 1 - \frac{(m-1) \sum \sigma_i^2 - \sum \sum r_{ik} \sigma_i \sigma_k}{(m-1) \left(\sum \sigma_i^2 + \sum \sum r_{ik} \sigma_i \sigma_k \right)} \\ &= \frac{(m-1) \sum \sigma_i^2 + (m-1) \sum \sum r_{ik} \sigma_i \sigma_k}{(m-1) S_T^2} \end{aligned}$$

$$\begin{aligned}
& - \frac{(m-1) \sum \sigma_i^2 - \sum \sum r_{jk} \sigma_j \sigma_k}{(m-1) S_r^2} \\
& = \frac{m \sum \sum r_{jk} \sigma_j \sigma_k + \sum \sum r_{jk} \sigma_j \sigma_k - \sum \sum r_{jk} \sigma_j \sigma_k}{(m-1) S_r^2} \\
& = \frac{m}{m-1} \left[\frac{\sum \sum r_{jk} \sigma_j \sigma_k}{S_r^2} \right] \quad (21)
\end{aligned}$$

one of the familiar forms of the KR-20 or alpha formula for reliability.

A conclusion to be drawn from these developments is that usually—i.e., when $N/(N-1)$ is nearly 1.0 and item difficulties differ—KR-20 should yield a somewhat larger estimate of reliability than KR-21. Actually, in many practical situations the difference between the two will be small—of the order of .05 or less. However, the fact of the difference between the two should be kept in mind when considering the standard error of measurement formulae developed in the next section.

Standard Error of Measurement Models

To arrive at an estimate of a standard error of measurement, one may solve equation (9) for $\sigma_{\text{pop-w}}^2$ in terms of r_{tt} , thus,

$$SEM(t) = \sqrt{\sigma_{\text{pop-w}}^2} = s_i \sqrt{1 - r_{tt}} \quad (22)$$

where, for convenience in further derivations, the small t may now be regarded as expressed in deviation-score form.² In this SEM one is, in effect, pooling the individual standard errors of measurement, σ_i , described in equation (1). Here, then, is an assumption of homogeneity of error variances. This is the same assumption as is involved in the SEM derivations based upon a consideration of the correlation between hypothetically parallel tests, where, however, it would be recognized as an assumption of homoscedasticity.

The SEM arrived at in (22) is the same as that which is obtained by means of what is called the true-score substitution model (Gulliksen, 1950). In this the obtained score is not regressed to give an estimate of the true score. The model for estimation is simply

$$\bar{t}_i = t_i \quad (23)$$

² Also for convenience t is still regarded as the total score (over m items) divided by m .

That is, the true score is estimated as the obtained score and the standard deviation for repeated observations of t_i is the *SEM* given in (22).

This model is different from a fallible-score substitution model in which an obtained score is regarded as an estimate of another obtained score. That is, in setting up confidence bounds around a given t_i score, the assumption (perhaps implicit) is that this score is to be substituted for others that might be obtained in another measurement of the same type. The estimation in this case can be symbolized as

$$l_{i2} = t_{i1} \quad (24)$$

where 1 and 2 indicate the first and second occasions respectively.

In this case the standard error for repeated observation is assumed to be the standard deviation of the difference between two obtained scores. Under the assumption that sigmas for separate tests (occasions) are equal (and equal to σ_t), this reduces thus

$$\begin{aligned} SEM(dif) &= \sqrt{\sigma_1^2 + \sigma_2^2 - 2r_{12}\sigma_1\sigma_2} = \sqrt{\sigma_t^2 + \sigma_t^2 - 2r_{12}\sigma_t^2} \\ &= \sigma_t \sqrt{2(1 - r_{12})} \end{aligned} \quad (25)$$

The r_{12} in this *SEM* is often regarded as different from the r_{tt} in equation (22), but the two can be seen to be the same in derivation and in most situations where an *SEM* is used. The r_{12} is the correlation between two fallible forms of a test. The r_{tt} , on the other hand, usually is viewed as the correlation between a given test and a hypothetical test just like it, as in the most widely used derivation of the KR-20 formula. That is, r_{tt} is defined as the correlation between two linear composites of the form

$$t_i = (x_{i1} + \cdots + x_{mi}) \quad (26)$$

$$t_i' = (x_{i1}' + \cdots + x_{mi}')$$

where t_i' is a hypothetical parallel measurement of t_i . Then

$$r_{tt} = \frac{\sum_i \sum_k r_{ih'} S_i S_h'}{\sqrt{\sum_i S_i^2 + \sum_i \sum_k r_{ik} S_i S_k} \sqrt{\sum_h S_h'^2 + \sum_h \sum_k r_{hk} S_h' S_k'}} \quad (27)$$

Stated in this way the problem is that the components involving primes are hypothetical and so (27) cannot be computed directly.

However, since it is assumed that the item variances and covariances in the hypothetical test are of the same magnitude as in the obtained test, the second term in the denominator (representing the variance of the hypothetical test) can be regarded as equal to the first term in the denominator and the average covariance between obtained and hypothetical components can be regarded as equal to the average covariance among the obtained components. That is, it can be assumed that

$$\sum S_k'^2 + \sum \sum r_{k'g} S_k' S_{g'} \cong \sum S_i'^2 + \sum \sum r_{ik} S_i' S_k' \quad (28)$$

and

$$\frac{\sum \sum r_{jk}' S_j S_k'}{m^2} = \frac{\sum \sum r_{jk} S_j S_k}{m(m-1)} \quad (29)$$

whence (28) and (29) can be solved for the hypothetical components on the left in terms of the obtainable components on the right and the results can be substituted into (27), to give the KR-20 formula derived in equation (21). By adding and subtracting $\sum S_j'^2$ in the numerator, this expression is put in the more common form.

$$r_{tt} = \frac{m}{m-1} \left(1 - \frac{\sum S_i'^2}{S_t'^2} \right) \quad (30)$$

If the assumption is made that items are dichotomous, so that $S_i'^2$ can be replaced by $p_i(1-p_i)$, where p_i is the eccentricity of item j , and it is further assumed that all p_i are equal, then KR-21 can be derived directly from (30). As noted earlier, however, it is not necessary to derive KR-21 in this way.

If, moreover, the standard deviations for the components in (21) are assumed to be equal and $\sum \sum r_{jk}$ is replaced by $m(m-1)\tilde{r}_{jk}$, where \tilde{r}_{jk} is an estimate of the "typical" correlation among components (perhaps the average), then (21) reduces to the general Spearman-Brown formula

$$r_{tt} = \frac{m\tilde{r}_{jk}}{1 + (m-1)\tilde{r}_{jk}} \quad (31)$$

as is well known. But now given that one knows the correlation between two fallible tests and this is the r_{12} given in equation (25) for the SEM, then this may be looked upon as one kind of \tilde{r}_{jk} estimate to use in equation (31). Moreover, since the standard error in this

case is to be estimated on, and applied to, just one of the two forms, the r_{ii} being calculated is for a test that is just *one* times as long as the tests that were correlated. Thus in this case m in (31) is 1 and

$$r_{ii} = \frac{(1)r_{12}}{1 + (1-1)r_{12}} = r_{12} \quad (32)$$

Under these assumptions, then, the true-score and fallible-score substitution models and the corresponding $SEM(t)$ and $SEM(dif)$ may be said to involve the same summary statistics (σ and r) and can thus be compared as to their implications in use.

In contrast to the substitution models are models in which the estimated score is obtained as a linear regression from the obtained score. The fallible-score regression model is the simplest of these. This is the usual fixed-variate regression model. One variable is estimated from another using the equation

$$\bar{t}_{i2} = r_{12} \left(\frac{\sigma_2}{\sigma_1} \right) t_{i1} \quad (33)$$

whence the standard error of estimate

$$\sigma(\bar{t} - t) = SEM(fr) = \sigma_t \sqrt{1 - r_{12}^2} \quad (34)$$

(fr stands for "fallible regression") is then treated as an SEM . The r_{12} in this case has the same meaning as the r_{12} introduced in the discussion of the fallible-score substitution model. Thus this SEM can be compared with the SEM of equation (22).

In a true-score regression model for error of measurement, the estimation equation is

$$\tau_i = r_{t\tau} \left(\frac{\sigma_\tau}{\sigma_t} \right) t_i \quad (35)$$

and the SEM is again, as in the case of the fallible-score regression model, the standard error of estimate

$$SEM(tr) = \sigma_\tau \sqrt{1 - r_{t\tau}^2} \quad (36)$$

(tr stands for true regression). But here the values involving τ , the true score, are hypothetical. Thus it is necessary to work out the computable form of these.

Using the assumptions of the classical true-score and error-score derivations, $r_{t\tau}^2$ can be seen to be equal to r_{tt} , as defined above, and the SEM can be seen to involve the same summary statistics as are

contained in the other *SEM*'s. That is, the fallible score, t_i , is assumed to be a simple linear combination of a true-score component and an error-score component. Then the correlation r_{tr} is derived as

$$r_{tr} = \frac{\sum t_r}{N\sigma_t\sigma_r} = \frac{\sum (\tau + e)\tau}{N\sigma_t\sigma_r} = \frac{\sigma_\tau^2 + r_{\tau e}\sigma_\tau\sigma_e}{\sigma_t\sigma_r} = \frac{\sigma_\tau}{\sigma_t} \quad (37)$$

where it is assumed that $r_{\tau e}$, the correlation between true-score and error-score components, is zero. Next the correlation between two fallible sets of measurements involving the same true-score component is obtained as an empirical estimate of reliability

$$\begin{aligned} r_{tt} &= \frac{\sum (\tau + e)(\tau + e)}{N\sigma_t\sigma_t} \\ &= \frac{\sigma_\tau^2 + r_{\tau e}\sigma_\tau\sigma_e + r_{e\tau}\sigma_\tau\sigma_e + r_{ee}\sigma_e\sigma_e}{\sigma_t\sigma_t} \\ &= \frac{\sigma_\tau^2}{\sigma_t^2} \end{aligned} \quad (38)$$

where it is assumed that $r_{\tau\tau}$, $r_{\tau e}$ and r_{ee} are zero and $\sigma_t = \sigma_{t'}$. Given these results, it is evident that r_{tt} in (38) is equal to r_{tr}^2 , the square of the value in (37), so $r_{tr} = \sqrt{r_{tt}}$. Also, (37) or (38) can be solved for σ_τ in terms of observables

$$\sigma_\tau = \sigma_t r_{tr} = \sigma_t \sqrt{r_{tt}} \quad (39)$$

When these values for the r_{tr} and σ_τ are substituted into equation (35), the estimation equation under the assumptions of the true-score regression model is seen to be

$$\begin{aligned} \tau_i &= \sqrt{r_{tt}} \left(\frac{\sigma_t \sqrt{r_{tt}}}{\sigma_t} \right) t_i \\ &= r_{tt} t_i \end{aligned} \quad (40)$$

(where it is assumed that the mean of the true scores equals the mean of the fallible scores). When the similar substitutions are made into equation (36), the standard error of measurement for the estimated true scores is seen to be

$$SEM(tr) = \sigma_t \sqrt{r_{tt}} \sqrt{1 - r_{tt}} \quad (41)$$

These developments provide a rather clear basis for comparisons of different estimates of a standard error of measurement and the associated confidence intervals for a given score. For they

suggest that the same σ and r_{tt} would enter into the calculation of each of the four *SEMs*.

Comparisons of the values of *SEMs* for a representative set of values of r_{tt} are shown in Table 1 for the standard score case (i.e. $\sigma = 1.0$). Here it can be seen that the order from smallest to largest *SEM* is

$$SEM(tr) < SEM(t) < SEM(fr) < SEM(dif)$$

and that *SEM(tr)* first increases and then decreases with increase in r_{tt} from .00 to 1.00, whereas all other *SEMs* decrease regularly with increase in r_{tt} .

The practical implications of these developments can be seen more concretely by considering the estimated scores and confidence boundaries that would be obtained for a given score when using a

TABLE 1
Comparisons of SEMs and Confidence Intervals

Reliability Value r_{tt}	Square of <i>SEM</i> Values for Different r_{tt} Values			
	<i>SEM(tr)</i> True-Score Regression $r_{tt}(1 - r_{tt})$	<i>SEM(t)</i> True-Score Substitution $(1 - r_{tt})$	<i>SEM(fr)</i> Fallible-Score Regression $(1 - r_{tt})$	<i>SEM(dif)</i> Fallible-Score Substitution $2(1 - r_{tt})$
1.0	.00	.00	.00	.00
.9	.09	.10	.19	.20
.8	.16	.20	.36	.40
.7	.21	.30	.51	.60
.6	.24	.40	.64	.80
.5	.25	.50	.75	1.00
.4	.24	.60	.84	1.20
.3	.21	.70	.91	1.40
.2	.16	.80	.96	1.60
.1	.09	.90	.99	1.80
.0	.00	1.00	1.00	2.00

*SEMs and Corresponding Confidence Intervals when
 r_{tt} is .91 and σ_t is 15*

<i>SEM</i> Value	Score (IQ)		.95 Confidence Bounds ± 2 <i>SEM</i>
	Obtained	Estimated	
<i>SEM(tr)</i> = 4.3	130	127.3	118.7 to 135.9
<i>SEM(t)</i> = 4.5	130	130.0	121 to 139
<i>SEM(fr)</i> = 6.2	130	127.3	114.9 to 139.7
<i>SEM(dif)</i> = 6.4	130	130.0	117.2 to 142.8

well known test. At the foot of Table 1 is shown a comparison of the intervals estimated for an IQ score of 130 obtained with a test having a reliability of .91 and a standard deviation of 15 IQ points. It can be seen that an applied worker, such as a teacher, might get rather different ideas about a person's IQ depending upon which information about confidence bounds was provided.

Of importance in the present context is the suggestion that different conclusions about the boundaries for an obtained score derive from the *same* computed summary statistics. This is contrary to what frequently seems to be assumed. In particular it seems to be assumed rather frequently that $SEM(fr)$ and $SEM(dif)$, in contrast to $SEM(tr)$ and $SEM(t)$, would be used under quite different conditions, involving quite different r coefficients. Here the suggestion is that the differences are apparent but not real and that what should dictate choice of one or the other SEM are the factors which would lead one to prefer one or the other model for variability of an obtained measurement. It does not follow that because one has a correlation, r_{12} , between two equivalent tests, he must use either $SEM(fr)$ or $SEM(dif)$, and that because he has a KR-20 or KR-21 coefficient, he must use either $SEM(tr)$ or $SEM(t)$. Given either of these kinds of coefficients, an investigator might use any one of the four SEM models.

It should be emphasized also that because the models for reliability (outlined in the first section of this article) lead to somewhat different values for an r_{tt} , still another source of difference in setting confidence bounds is introduced. On the same data, for example, the r_{tt} computed by means of equation (9) (the KR-21 Lord model) leads to a value of .90, whereas the r_{tt} computed by means of equation (21) (the KR-20 Hoyt model) gives a value of .94. These differences will produce differences in any SEM computed on the basis of an estimate of r_{tt} .

These technical issues concerning the way error is to be estimated for a given set of observations do not in any way dispute the fact that different kinds of variability can be represented as "error."³ Variability between scores obtained with tests that can

³ As the botanist regards a weed as "a flower out of place," so the psychometrist can regard error as "variation out of place," thus emphasizing that what is regarded as error at one time might not be so regarded at another time.

be regarded as replacements for each other (as in computing equivalency reliability), variability between measurements obtained at quite different times (as in computing stability reliability), variability among responses obtained on a single occasion (as in obtaining internal consistency reliability) and variability between observers (as in computing conspect or inter-rater reliability) may be designated as "error" in any one of the formulae for reliability or *SEM*. The concepts of error represented by these ways of regarding "variability out of place" have no necessary relationship to the mathematical theory outlined here. The mathematical theory may be thought of as a metatheory which can be employed in conjunction with any one of the four (and perhaps other) theories stipulating the kind of behavioral variability that is regarded as error. For example whether the basic observations are test-retest scores obtained over a ten year interval or separate measures of thirst obtained on a given occasion, one can use either of the two r_{tt} 's or any one of the four *SEM*s discussed in this paper.

REFERENCES

- Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-334.
- Cronbach, L. J., Rajaratnam, N., and Gleser, G. C. Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 1963, 16, 137-163.
- Gulliksen, H. Theory of Mental Tests. New York: Wiley, 1950.
- Henrysson, S. Different measures of reliability. Research Bulletin, Institute of Educational Research, Teachers College, Stockholm, Sweden, 1959.
- Hoyt, C. Test reliability obtained by analysis of variance. *Psychometrika*, 1941, 6, 153-160.
- Horn, J. L. Equations representing combinations of components in scoring psychological variables. *Acta Psychologica*, 1963, 21, 184-217.
- Horst, P. Psychological measurement and prediction. Belmont, California: Wadsworth, 1966.
- Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, 2, 151-160.
- Lord, F. M. Estimating test reliability. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1955, 15, 325-336. (a)
- Lord, F. M. Sampling fluctuations resulting from the sampling of test items. *Psychometrika*, 1955, 20, 1-22. (b)
- Lord, F. M. Do tests of the same length have the same standard errors of measurement? EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1957, 17, 510-521.

- Lord, F. M. Statistical inferences about true scores. *Psychometrika*, 1959, 24, 1-18. (a)
- Lord, F. M. An approach to mental test theory. *Psychometrika*, 1959, 24, 282-302. (b)
- Lord, F. M. Test reliability—A correction. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1962, 22, 511-512.
- McNemar, Q. The revision of the Stanford-Binet scale. Boston: Houghton Mifflin, 1942.
- Novick, M. R. The axioms and principal results of classical test theory. *Journal Mathematical Psychology*, 1966, 3, 1-18.
- Novick, M. R. and Lewis, C. Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 1967, 32, 1-13.
- Penfield, D. A. An empirical investigation of the approximate sampling distribution of Kuder-Richardson Twenty. PhD Dissertation, University of California, Berkeley, 1967.
- Sheffé, H. A. The analysis of variance. New York: Wiley, 1959.
- Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

A MEASURE OF AGREEMENT AMONG SUBJECTIVE JUDGMENTS

K. H. LU

Department of Biostatistics
University of Oregon Dental School

In many areas of inquiries, the variable of interest often does not lend itself to direct physical measurement; its mensuration, therefore must rely upon the subjective judgments of the observer. For example, the evaluation of works in the arts and letters, the weighing of certain experimental and clinical data in biomedical research, psychology, sociology, and education are all but a few of the areas of this kind. Due to the multitude of factors associated with the variable, precise and detailed considerations of all of these factors are either impractical or impalable. Consequently, the common practice has been that a set of criteria is set up beforehand such that the variation of the variable may be classified into a set of ordered categories. Suppose we imagine a situation involving n subjects, m judges, and t categories. Each of the n subjects is examined and assigned into one of the t categories independently by each of the m judges. Our task is to devise a metric to measure the intensity of agreement among the m judges.

At first glance, the $m \times n$ observations may be grouped into a $m \times t$ contingency table with each row summing up to n , thus we have a situation where the chi square test of homogeneity among the m judges with respect to their relative frequencies in the t categories might serve as a means of measuring agreement. On closer examination, however, one recognizes that the chi square test will not suffice because (1) a significant chi square value simply signifies that some or all of the m judges are significantly different from each other, but says nothing about the agreement among them, and (2) an insignificant chi square merely denotes the lack of evidence

against the hypothesis that the relative frequencies of the m judges with respect to the t categories are the same, it again says nothing about how closely they resemble each other, namely, the intensity of agreement. To take another view, the above mentioned situation may be viewed as a multi-judge rank correlation problem which allows ties in the t ranks. The rank correlation viewpoint, although theoretically acceptable, in actual application, the frequent occurrences of tied ranks renders the resultant rank correlation cumbersome to calculate and somewhat powerless in meaning (Kendall 1955).

The purposes of this paper are twofold: (1) To describe a weighing procedure of the categories, i.e., to assign a value to each category based on a transformation from the data's own distribution, and (2) to devise a coefficient of agreement of the judges calculated from the transforms of the categorical data.

Theoretical Consideration

Let a set of n subjects (or products) be judged by a set of m judges with respect to some attribute X , according to some descriptive criterion. Assume that attribute X is conceptually measurable on a continuous scale, but due to practical limitations, direct physical measuring metrics are not possible; instead, it can only be rated in terms of an order set of nonoverlapping categories as described by the criterion. The set of categories may consist of, say, t categories, $C_1, C_2 \dots C_k \dots C_t$, such that $X_i < X_j$ if $X_i \in C_i, X_j \in C_j$, and $C_i < C_j$ in terms of the continuum of X . The observed results of the above may be represented by the following array:

	J_1	$J_2 \dots J_j \dots J_m$
S_1	X_{11}	$X_{12} \quad X_{1j} \quad X_{1m}$
S_2	X_{21}	$X_{22} \quad X_{2j} \quad X_{2m}$
\vdots		
S_i	X_{i1}	$X_{i2} \quad X_{ij} \quad X_{im}$
\vdots		
S_n	X_{n1}	$X_{n2} \quad X_{nj} \quad X_{nm}$

Where X_{ij} is the judgment of the i th subject by the j th judge and may assume the "values" of C_k ($i = 1, 2, \dots n; j = 1, 2, \dots m$; and $k = 1, 2, \dots t$). In order to devise a metric to measure the

"degree" of agreement among the m judges, it is necessary that we now delineate the concept of agreement and define the meaning of its measure.

The simplest case of agreement may be illustrated by the case where two judges placing a single subject into one of two categories A and B , such that $A < B$, say. Clearly the events AA and BB constitute agreement and the events AB and BA the disagreement. However, when there are more than two judges and two categories involved, the demarcations between agreement and disagreement are no longer so clear-cut because the term agreement begins to assume a meaning of gradation in a quantitative sense. Suppose there are three judges placing a subject into one of three categories A , B , and C , and where $A < B < C$. Obviously, the events AAA , BBB , and CCC still constitute the agreement. But AAB , is certainly in closer agreement than AAC , because B is closer to A than C . However, can we assume that AAB is in the same degree of agreement as BBC ? Is ABC in worse agreement than AAC ? In view of all these unsettling points, what then constitutes the lower end of agreement? The situation becomes more critical when there are more judges than categories, (i.e., $m > t$), because then agreement between certain judges is inevitable. It is clear then that while we have no problem in defining the maximum agreement among m judges in terms of t categories ($m, t > 2$), we are facing a dilemma in defining the other extreme of the agreement scale, namely, the disagreement, or the minimum agreement. In order to bring the above dilemma into some manageable fashion, it becomes necessary to define the minimum agreement in terms of some clearly defined criteria.

While there are many different ways to define the minimum agreement, we shall only choose two of them; namely, (1) the maximum within subject variance, and (2) the maximum within subject entropy. As the names implied, the first measure is a metric of agreement from the analysis of variance point of view and the second is from the information theory point of view. Suppose that appropriate weights y_k may be assigned to categories C_k respectively. Let S_i^2 = variances of ratings of the i th subject, we have

$$S_i^2 = \frac{\sum_{j=1}^m (y_{ij} - \bar{y}_i)^2}{m - 1}$$

and let H_i = entropy of ratings of the i th subject.

$$= - \sum_{k=1}^t P_{ik} \log P_{ik}$$

where

$$P_{ik} = \frac{\text{no. of judges rated the } i\text{th subject as } C_k}{m}, \quad (\text{Lu } 1968).$$

We see in the case of perfect agreement, all m ratings are in a particular C_k , hence all have identical y_{ik} , then

$$S_i^2 = 0, \quad H_i = 0$$

The maximum S_i^2 will occur if the m ratings are distributed $(m/2)C_1$ and $(m/2)C_t$, because since

$$\begin{aligned} \bar{y} &= \frac{\frac{m}{2}(y_1 + y_t)}{m} = \frac{y_1 + y_t}{2}, \\ S_i^2 &= \frac{\frac{m}{2}(y_1 - \bar{y})^2 + \frac{m}{2}(y_t - \bar{y})^2}{m - 1} \\ &= \frac{m}{4(m - 1)}(y_1 - y_t)^2 \end{aligned}$$

is a maximum since $(y_1 - y_t)^2 > (y_i - y_r)^2$ for $i > 1$ and $r < t$. From the analysis of variance point of view this is the minimum agreement, i.e., the m judges have divided into two camps of opinions as farther apart as possible. The maximum entropy will occur if $P_i = 1/t$ for all i . In other words, the m judges assign any one of the t ratings to the i th subject with equal likelihood. From this point of view, this is the worst the judges can do, hence any agreement among the judges is at random, therefore, the agreement is not meaningful, and consequently, the minimum agreement.

It is now abundantly clear that the minimum agreement as defined by the maximum variance S_i^2 is not the minimum agreement defined by the maximum entropy H_i . To put it in words, while

$$S_i^2 = \frac{m}{4(m - 1)}(y_1 - y_t)^2$$

may be a maximum variance, but its entropy H_i is certainly not a maximum. A fact easily discernable is that there is perfect agree-

ment among each set of $m/2$ judges. On the other hand, the random assignment of rating to the i th subject may yield a maximum entropy H_i , but the within subject variance is certainly not the largest.

Thus, the variance measure S_i^2 cannot account for the agreement within sets of agreeing judges, the entropy measure cannot distinguish the difference between magnitude of agreement. For instance, suppose we have two situations, where the m ratings are

$$(1) \quad \frac{m}{2} y_1 \quad \text{and} \quad \frac{m}{2} y_i, \quad \text{and} \quad (2) \quad \frac{m}{2} y_2 \quad \text{and} \quad \frac{m}{2} y_{i-1}$$

since $y_1 < y_2 \cdots < y_{i-1} < y_i$, the variance measure will show the first case having less agreement than the second, but the entropy measures of the two cases will be identical. It would be desirable for the devised measure of agreement to utilize both the variance and the information view points.

Since in reality, one seldom has maximum or minimum agreement, and if agreement is present and detectable at all, it must lie somewhere between the maximum and the random agreement. We shall, therefore, use the within variance under conditions of maximum entropy. Thus, the coefficient of agreement, A , may be defined as follows:

$$A = \frac{\sigma_H^2 - S_i^2}{\sigma_H^2} \dots 1$$

where

S_i^2 = the observed within subject variance.

σ_H^2 = expected within subject variance under conditions of maximum entropy, i.e., all ratings are equally likely

$$Pr \{X_{ii}\} = y_k = \frac{1}{t} = \sum_{k=1}^t \frac{y_k^2}{t} - \left(\sum_{k=1}^t \frac{y_k}{t} \right)^2.$$

We see that

$$A \rightarrow 0 \quad \text{as} \quad S_i^2 \rightarrow \sigma_H^2$$

$$A \rightarrow 1 \quad \text{as} \quad S_i^2 \rightarrow 0.$$

A can never be of indeterminate form because $\sigma_H^2 > 0$. One point needs to be stressed here is that $A < 0$ as $S_i^2 > \sigma_H^2$. In such case, it implies that the judges agree less among themselves than random,

i.e., more disagreement than agreement. Our present interest is to measure agreement, consequently, we do not entertain any $S_i^2 > \sigma_H^2$. If it should occur, a coefficient of disagreement may be defined as

$$D = \frac{S_m^2 - S_i^2}{S_m^2}$$

where $S_i^2 > S_H^2$ and

$$S_m^2 = \frac{m}{4(m-1)} (y_1 - y_t)^2$$

It is quite clear that the feasibility of computing the coefficient of agreement depends solely upon our ability to obtain the appropriate weights y_k for each of the categories C_k . In the passages to follow we shall show such a procedure.

*The Determination of Appropriate Weights
for the t Categories*

Let n subjects be judged by m judges according to some criteria into t categories. Since the range of variation of agreement is defined in terms of intra-subject variance, our next task is to assign a set of meaningful weights to each of the C_k , such that the variances may be calculated. Let us regroup the $m \times n$ array as follows:

	C_1	C_2	\dots	C_k	\dots	C_t	Σ
J_1	n_{11}	n_{12}		n_{1k}		n_{1t}	n
J_2	n_{21}	n_{22}		n_{2k}		n_{2t}	n
\vdots							
J_i	n_{i1}	n_{i2}		n_{ik}		n_{it}	n
\vdots							
J_m	n_{m1}	n_{m2}		n_{mk}		n_{mt}	n
Σ	$n_{1.}$	$n_{2.}$		$n_{k.}$		$n_{t.}$	mn

where n_{jk} is the count of individuals placed in the k th category by the j th judge.

Let $f(x)$ be the frequency function of random variable X defined in terms of C_k and assume the first two moments of X exist. We see that each C_k constitutes a segment of the area under the curve as follows:

$$Pr \{X \in C_k\} = \int_{c_k} f(x) dx.$$

Now let random variable y be defined as the distribution function of X , say $F(x)$ such that

$$y = F(x) = \int_{-\infty}^x f(u) du,$$

then

$$\frac{dy}{dx} = f(x), \quad \text{and} \quad \frac{dx}{dy} = \frac{1}{f(x)}$$

The frequency function of y is therefore

$$\begin{aligned} g(y) &= f(x) \left| \frac{dx}{dy} \right| dy. \\ &= \frac{f(x)}{f(x)} dy \end{aligned}$$

We see the transformation $x \rightarrow y$ has the unique property that regardless what the form $f(x)$ might have been, the transformed variable y is always a variable following the uniform distribution $g(y)$ with $\mu = 1/2$ and $\sigma^2 = 1/12$.

In practice, we let $P_r = n_r/mn$, and

$$y_k = F(x) = \sum_{r=1}^{k-1} P_r + \frac{1}{2} P_k \quad k = 1, 2, \dots, t.$$

It can be shown that

$$E(y) = \frac{1}{2} \quad \text{and} \quad E(S_y^2) = \frac{1}{12}.$$

We have thus obtained a set of values for the t categories. The transformed variable y is of a probabilistic scale in terms of the distribution function of X (Bross 1958). Substituting the computed y_k 's for the correspondent C_k 's for each x_{ij} , we shall show an array of $m \times n$ entries

	J_1	J_2	J_i	J_m
S_1	X_{11}	X_{12}	X_{1i}	X_{1m}
S_2	X_{21}	X_{22}	X_{2i}	X_{2m}
S_i	X_{i1}	X_{i2}	X_{ii}	X_{im}
S_n	X_{n1}	X_{n2}	X_{ni}	X_{nm}
$i = 1, 2, \dots, m$			$j = 1, 2, \dots, m$	

$$X_{ij} = y_1, y_2 \dots y_k \dots y_t$$

Thus the above array may be considered as a two-way factorial design (subjects \times judges) with m and n levels respectively.

An analysis of variance may be performed as follows:

	df	m.s.	m.s. is estimate
Between subjects	$n - 1$	S_B^2	$\sigma_i^2 + m\sigma_s^2$
Within subjects	$n(m - 1)$	S_i^2	σ_i^2
Total	$nm - 1$		

The within subject variance under maximum entropy condition is by definition

$$\sigma_H^2 = \frac{1}{t} \sum y_k^2 - \left(\frac{\sum y_k}{t} \right)^2 \dots 2$$

A test of significance of A may be conducted indirectly. We shall choose to test the hypothesis that the assignments of subjects to the categories by the judges are at random, i.e., equally likely for all t categories. Then we would expect that

$$E(S_i^2) = \sigma_H^2.$$

Thus the statistic

$$\theta = \frac{S_i^2}{\sigma_H^2}$$

is χ^2/df distributed with $n(m - 1)$ degrees of freedom. If we reject the hypothesis that $E(S_i^2) = \sigma_H^2$ we can conclude that A is significantly different from zero (Dixon and Massey 1957).

Substituting the necessary quantities S_i^2 and σ_H^2 into equation (1), we have the coefficient of agreement

$$A = \frac{\sigma_H^2 - S_i^2}{\sigma_H^2}$$

A remark concerning the appropriateness of the above analysis of variance operation would seem in order here. One might have noticed that the transformed variable y is uniformly distributed, whereas in analysis of variance, normality of the variable is required. The violation of normality is not unique with the present case in fact, the use of non-normal data as if they were normal are quite common in statistical works in education and psychology. For instance, the computation of the coefficient of reliability is based on test scores of a dichotomous nature. The valid argument for

tolerating this violation is not that with company one has strength for wrong doing, instead, it is due to the robustness of the analysis of variance technique, the effect of non-normality is negligible as long as the subclass numbers are equal (Tukey, 1956, 57; Scheffe, 1964). The two-way factorial table in our present case certainly has this property.

Illustrative Examples

Suppose there are 12 subjects who each has performed a task. As shown in Table 1 each of the 12 subjects was rated by two sets of four judges each. Each subject was assigned a grade, say A, B, C, D, and F.

TABLE 1
Judge Sets

	I				II			
	R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4
1	F	F	F	F	F	F	F	F
2	F	F	D	F	D	F	F	F
3	F	D	D	F	D	D	F	F
4	D	D	D	D*	C	D	D	D
5	D	D	D	D*	C	D	C	D
6	D	D	C	D	C	C	C	D
7	C	C	C	C*	C	C	B	D
8	C	C	C	C*	B	B	B	C
9	C	B	B	B*	B	B	A	C
10	B	B	B	B*	B	A	A	B
11	A	B	B	A	A	A	A	B*
12	A	A	A	A	A	A	A	A

* The better agreement of the two sets. The measure of agreement A should show a higher value for Set I than for Set II.

The computation of weights and computation of S_i^2

SET I

	1	2	3	4	5
	$n_{i.}$	$\frac{1}{2}n_{i.}$	$\sum_{i=1}^{i-1} n_{i.}$	$(2) + (3)$	$y_i = (4)/n$
1(F)	9	4.5	0	4.5	.09375
2(D)	14	7.0	9	16.0	.33333
3(C)	10	5.0	23	28.0	.58333
4(B)	9	4.5	33	37.5	.78125
5(A)	6	3.0	42	45.0	.93750

SET II

1(F)	9	4.5	0	4.5	.09375
2(D)	10	5.0	9	14.0	.29107
3(C)	10	5.0	19	24.0	.50000
4(B)	9	4.5	29	33.5	.69792
5(A)	10	5.0	38	43.0	.89583

Substituting the calculated y_i 's into the original table, we compute the analysis of variance for each set.

	<i>df</i>	<i>ms</i> (I)	<i>ms</i> (II)
Subject	11	.32753	.29089
Within	36	.00543	.01774

$$\sigma_H^2(I) = \frac{1}{5} \sum y_i^2 - (\frac{1}{5} \sum y_i)^2 = .38907 - .29793 = .09194$$

$$\sigma_H^2(II) = \frac{1}{5} \sum y_i^2 - (\frac{1}{5} \sum y_i)^2 = .32609 - .24573 = .08036$$

Tests of significance

$$\theta_1 = \frac{.00543}{.09194} = .05906^{**} \quad df = 36$$

$$\theta_2 = \frac{.01774}{.08036} = .22075^{**} \quad df = 36$$

The coefficients of agreement are:

$$A_I = \frac{.09194 - .00543}{.09194} = .94094$$

$$A_{II} = \frac{.08036 - .01774}{.08036} = .77924$$

REFERENCES

- Bross, I. D. J. How to use Redit analysis; *Biometrics*, 1958, 14, 18-38.
- Dixon, W. J. and F. J. Massey, Jr. *Introductions to Statistical Analysis*. McGraw-Hill, New York, 1957.
- Kendall, M. G. *Rank correlation methods*. Hafner, New York, 1955.
- Lu, K. H. An information and discriminant analysis of fingerprint patterns pertaining to identification of mongolism and mental retardation; *American Journal of Human Genetics*, 1968, 20, 24-43.
- Scheffe, Henry. *The analysis of variance*. John Wiley & Sons, Inc., New York, 1964.

THE INTERPRETATION OF REGRESSION COEFFICIENTS IN A SCHOOL EFFECTS MODEL¹

ROBERT L. LINN AND CHARLES E. WERTS

Educational Testing Service

LEDYARD R. TUCKER

University of Illinois

LARGE scale multipurpose "shotgun" surveys in research concerned with school effects have generally been used to obtain descriptive information about schools and students and to locate student background and school characteristics associated with student growth during the school years. The intense and often irreconcilable debate about the interpretation of Coleman's (1966) monumental "Equality of Educational Opportunity" reflects the highly speculative nature of any inferences about the causes of change or growth in students. In light of these hazards it is perhaps premature to consider the further refinement of using survey data to construct a model simulating the effects of schools on students, e.g., allowing for such questions as: What would happen to reading skills if we increased the number, variety, and quality of books available to students by a specified amount? Even if a reasonable simulation model cannot be constructed now, we believe that the process of trying to construct it can be useful in several ways (Blalock and Blalock, 1968) notably:

1. In clarifying the numerous debates about which statistical procedure should be used. One cannot specify which statistics are appropriate until the (postulated) nature of the phenomena under study has been specified. Thus attempts to specify a logical model

¹ The research reported herein was performed pursuant to Grant No. OEG-1-6-061830-0650 Project No. 6-1830 with the Office of Education, U. S. Department of Health, Education, and Welfare.

will require spelling out the alternate hypotheses in ways which will, for example, suggest whether changes in mean, variation, and/or relative rank are relevant. Many educational researchers have implicitly equated correlation with "influence" without considering whether the particular influence will be reflected in a correlation.

2. In clarifying which data need to be collected in the *next* survey to help choose among the plausible alternative hypotheses for observed associations. It is a common experience for researchers who work out a series of alternate hypotheses to find that the data from previously collected "shotgun" surveys are not detailed enough in that particular area to provide any test between the alternatives.

3. In specifying the values and goals of education. As has been repeatedly noted, much survey research is a post mortem on past models rather than a consideration of the potentialities for change. Thus, if only a few schools have a progressive program in a given area, when all schools are analyzed together, the small percentage of variance accounted for by the "school effect" may reflect mainly the mediocrity of the past rather than the potential for the future. A widely used method with only a slight impact may account for a larger percentage of variance than a very potent but rarely used technique. Consideration of goals may also help avoid the common mistake of simultaneously analyzing on the same output scale schools with very disparate goals. This error confounds exposure with quality of curriculum offered, thus hindering the possible discovery of methods for more efficiently accomplishing the particular goals set by the school administrators. Furthermore, this process should lead to greater specification and discussion of alternative strategies for attaining the specified goals and the costs of each.

In this paper we have attempted to illustrate the concrete and detailed mode of thinking required in the above process, by considering how one might interpret regression coefficients in relation to several alternate strategies for allocating resources. While the real world certainly does not really operate according to the model discussed, at least predictions can be made which further research can either confirm or disprove. Optimistically, the result of this type of concrete consideration may be a more "reasonably specific" formulation of what we do not know in a region of inquiry, of why we want to know it, and, in the favorable case, of how we might

proceed to find out" (Merton, 1968). Our commitment is to the "strong inference" approach enunciated by Platt (1964).

A Regression Model

It is customary to identify three classes of variables in school effects studies: measures of student "input," measures of student "output," and measures of the "school characteristics." For illustrative purposes we shall use an example with a single input variable X , an output variable Y , and a school variable W . A linear regression model for this case would consist of:

$$\hat{Y}_{ij} = B_w W_j + B_x X_{ij} + A$$

$$e_{ij} = Y_{ij} - \hat{Y}_{ij}$$

where

X_{ij} is the input measure for student i in school j ,

Y_{ij} is the output measure for student i in school j ,

W_j is the "characteristic" measure for school j , and A is a constant.

When the covariances of e with w (i.e., C_{ew}) and e with x (i.e., C_{ex}) are assumed to be zero, least squares analysis can be used to find B_w , B_x , and A :

$$B_w = \frac{S_x^2 C_{wy} - C_{xw} C_{xy}}{S_x^2 S_w^2 - C_{xw}^2},$$

$$B_x = \frac{S_w^2 C_{xy} - C_{xw} C_{wy}}{S_x^2 S_w^2 - C_{xw}^2},$$

$$A = \bar{Y} - B_w \bar{W} - B_x \bar{X},$$

where

\bar{X} , \bar{W} and \bar{Y} are the means of X , W and Y , respectively, S_x^2 , S_w^2 , S_y^2 are the variances of X , W and Y , respectively, and C_{xy} , C_{wy} , and C_{wx} are the covariances between X and Y , W and Y , and W and X , respectively.

The variance of Y is

$$S_y^2 = S_{\hat{y}}^2 + S_e^2,$$

where

$$S_{\hat{y}}^2 = C_{wy} B_w + C_{xy} B_x,$$

and S_e^2 is the variance of e .

If a regression model is used for simulation purposes it is important to remember the kinds of theoretical assumptions being made:

1. Most importantly, that a linear additive model with equal unit scale assumptions is at least plausible in light of what is known about the effect to be measured. Does reality operate on the least squares principle?

2. That all input and school variables influencing the output are in the model equation(s). Within a linear model, the absence of relevant influences will, in general, lead to "specification error" i.e., bias in calculated parameters due to incorrect specification of the structural model. If, however, an influence on the output is uncorrelated with all independent variables specified in the equations, its absence will, in general, not bias estimates of the weights for the variables included. The error term in a regression model represents all those unmeasured implicit factors which influence output, but are assumed or known to be uncorrelated with the independent variables in the equation(s). For a more complete discussion of specification error see Theil (1957).

3. All independent variables are measured without error or appropriate corrections are introduced for such errors, e.g., corrections for unreliability. Random measurement error in the output variable should not bias regression estimates. Cochran (1968) reviews the effects of and corrections for errors of measurement. Linear regression models are considered in sections eight thru eleven of Cochran's review.

4. If the correlations among independent variables become very high and/or multiple forms of the same independent variable are entered into the analysis in different guises, not only will the statistical problems of multicollinearity (Farrar and Glauber, 1967) be involved, but reasonable substantive interpretation of such results may become almost impossible (Gordon, 1968).

5. That unstandardized regression weights do not, in general, lead to estimates of the relative "importance" of effects but may, nonetheless, be more suitable for making inferences about "influences" in a linear model. Blalock (1967) and Tukey (1954) have argued that the unstandardized regression weights are to be preferred to standardized weights or correlations in attempting to investigate influences primarily because of their relative independence of sample characteristics.

The requisite theoretical and statistical assumptions will usually be such as to render any interpretation of calculated regression weights quite tentative. Unless the study has been designed to elaborate the relationship of this model to reality, i.e., to eliminate alternate hypotheses about the nature of reality, the calculated regression weights may have no useful interpretation.

Interpretation of Change

Suppose that the school characteristic could be manipulated while keeping the distribution of the student input measure constant. Let there be two situations, α and β , corresponding to two distributions of W , i.e., two sets of values of the school characteristic measure for each school, $W_{j\alpha}$ and $W_{j\beta}$.

Define Z as the difference between the two values of the school atmosphere:

$$Z_i = W_{i\beta} - W_{i\alpha} \quad (1)$$

or

$$W_{i\beta} = W_{i\alpha} + Z_i.$$

Suppose, further, that A , B_w , B_x , and S_e^2 are constant for the two situations. Let $S_{w\alpha}^2$ and $S_{w\beta}^2$ be the variances of W for the two situations; $C_{zw\alpha}$ and $C_{zw\beta}$ be the covariances between Z and W for the two situations; $C_{xw\alpha}$ and $C_{xw\beta}$ be the covariances between X and W for the two situations; and $C_{yw\alpha}$ and $C_{yw\beta}$ be the covariances between Y and W for the two situations.

The preceding assumes that there will be different distributions of Y for the two situations, Y_α and Y_β , such that

$$\hat{Y}_{i\alpha} = B_w W_{i\alpha} + B_x X_{i\alpha} + A, \quad (2)$$

and

$$\hat{Y}_{i\beta} = B_w W_{i\beta} + B_x X_{i\beta} + A. \quad (3)$$

It follows directly from (1), (2), and (3) that

$$\hat{Y}_{i\beta} - \hat{Y}_{i\alpha} = B_w Z_i.$$

Thus, B_w is a measure of the effect of Z on the change in predicted output for any given individual student in school j . Further, if it is assumed that the mean error is zero in both situations, i.e., $\bar{e}_\alpha = \bar{e}_\beta = 0$ then B_w can be interpreted as a measure of the effect of Z on the change in means on Y , since

$$\bar{Y}_\beta - \bar{Y}_\alpha = B_w \bar{Z}, \quad (4)$$

which when $\bar{W}_\beta \neq \bar{W}_\alpha$ yields

$$B_w = \frac{\bar{Y}_\beta - \bar{Y}_\alpha}{\bar{W}_\beta - \bar{W}_\alpha}.$$

In general then, the regression weight for W can, in this model, be interpreted as the unit change in output (Y) that can be expected from a unit change in the school atmosphere (W), if input and implicit factors (e) remain constant. The regression weight for X indicates the unit change in output (Y) that can be expected from a unit change in the input (X), if the school characteristic and the implicit factors (e) remain constant. If both input and the school characteristic change one unit (e constant) then Y will change $B_w + B_x$ units. If input influences the school environment and all other influences on school environment are independent of input, i.e., $W = B_{wx}X + e'$ then a change of one unit in X will have two effects: (a) to change the environment B_{wx} units which in turn result in $(B_{wx}B_w)$ units of change in output, and (b) X will, in addition, directly change output B_x units. When the school atmosphere is a group composition effect which is necessarily influenced by input, it is unreasonable to think of changing the school variable and maintaining input constant. It follows that no interpretation of regression coefficients can proceed apart from a consideration of the nature of the effect being studied and the network or context in which the effect occurs.

Further understanding of regression coefficients can be obtained by examining the relationship of B_w to changes in output variance. It can be shown that under the assumptions specified above the change in the variance of Y can be expressed as:

$$S_{y\beta}^2 - S_{y\alpha}^2 = 2C_{zx}B_wB_x + (S_{w\beta}^2 - S_{w\alpha}^2)B_w^2. \quad (5)$$

Consider three possible strategies for allocating resources (assuming high X is "better" and higher W is more "desirable" in terms of output) to improve the school environment W : (1) "random allocation" of resources so that the level of student input is uncorrelated with the change in school environment ($C_{zx} = 0$); (2) the "rich get richer" approach in which the better the students entering a school, the more resources are provided to improve the school characteristic (i.e., $r_{zx} = +1.0$); and (3) the "compensatory education model" in which the worse the student entering a school

the more resources are provided (i.e., $r_{\alpha\alpha} = -1.0$). For case one, the "random allocation" strategy:

$$\frac{S_{y\beta}^2 - S_{y\alpha}^2}{S_{w\beta}^2 - S_{w\alpha}^2} = B_w^2, \quad (6)$$

provided that $S_{w\beta}^2 \neq S_{w\alpha}^2$.

Thus, in this case the square of the regression coefficient for the measure of the school characteristic is equal to the ratio of the difference between the variances of the output for the two conditions to the difference between the variances in the school characteristic measures for the two conditions. For case two, the "rich get richer" strategy, the difference between the variances of the output for the two conditions becomes

$$S_{y\beta}^2 - S_{y\alpha}^2 = 2S_z S_x B_w B_x + (S_{w\beta}^2 - S_{w\alpha}^2) B_w^2, \quad (7)$$

since $C_{zx} = S_z S_x$. Similarly, for case three the compensatory education model:

$$S_{y\beta}^2 - S_{y\alpha}^2 = -2S_z S_x B_w B_x + (S_{w\beta}^2 - S_{w\alpha}^2) B_w^2 \quad (8)$$

since $C_{zx} = -S_z S_x$.

A simple illustration may help clarify some of the implications of equations (4) thru (8). Suppose X and Y were scores on a physics test at the beginning and at the end of a second semester of physics, and that W_j was the expenditure per student for the second semester of physics at school j . Now suppose that the following regression equation was obtained:

$$\hat{Y}_{ij\alpha} = .5W_{i\alpha} + .8X_{ij} + 0.$$

Given the assumptions of the model, a change in expenditure per student from $W_{j\alpha}$ to $W_{j\beta}$ would result in a predicted score for a student at school j of

$$\begin{aligned} \hat{Y}_{ij\beta} &= \hat{Y}_{ij\alpha} + .5(W_{j\beta} - W_{j\alpha}), \\ &= \hat{Y}_{ij\alpha} + .5(Z_j). \end{aligned}$$

If a change in state allocations resulted in $Z_j = K$ for all j , then:

$$\bar{Y}_\beta - \bar{Y}_\alpha = .5K,$$

and $S_{y\beta}^2 - S_{y\alpha}^2 = 0$. On the other hand, if there was no mean change in allocations and, therefore, expenditures (i.e., $\bar{Z} = 0$) then

$$\bar{Y}_\beta - \bar{Y}_\alpha = 0.$$

Adjustments in the allocations to individual schools, while maintaining $\bar{Z} = 0$, would have quite different implications for the variance of the output scores depending on the correlation between Z and X . If adjustments in W were made that were completely unrelated to X as in the "random allocation" strategy, then the change in variance of Y would depend entirely on the change in the variance of W , i.e.,

$S_{y\beta}^2 - S_{y\alpha}^2 = .25 (S_{w\beta}^2 - S_{w\alpha}^2)$, which, in turn, could be expressed in terms of the correlation between Z and W , $r_{zw\alpha}$:

$$S_{y\beta}^2 - S_{y\alpha}^2 = .25S_z(2r_{zw\alpha}S_{w\alpha} + S_z).$$

Thus, for case one when B_w is positive $S_{y\beta}^2$ will be larger than $S_{y\alpha}^2$ if $r_{zw\alpha}$ is positive. If $r_{zw\alpha}$ is negative then $S_{y\beta}^2$ may be larger or smaller than $S_{y\alpha}^2$ depending on the relative magnitudes of $S_{w\alpha}$ and S_z .

If adjustments for the illustration were made such that allocations were increased at schools with the initially best students and decreased at schools with the initially poorest students as in the "rich get richer" strategy (i.e., $r_{zx} > 0$) then:

$$\begin{aligned} S_{y\beta}^2 - S_{y\alpha}^2 &= .8C_{zx} + .25(S_{w\beta}^2 - S_{w\alpha}^2), \\ &= S_z(.8r_{zx}S_x + .5r_{zw\alpha}S_{w\alpha} + .25S_z). \end{aligned}$$

At the extreme where $r_{zx} = 1$ (i.e., case two)

$$S_{y\beta}^2 - S_{y\alpha}^2 = S_z(.8S_x + .5r_{zw\alpha}S_{w\alpha} + .25S_z),$$

which will, in general, be positive except when $r_{zw\alpha}$ is negative and $S_{w\alpha}$ is large relative to S_w and S_z . In the "compensatory education model" the allocations are increased at the schools with the initially poorest students and decreased at the schools with the initially best students. The difference between the output variances for the illustration for the most extreme compensatory education model (i.e., $r_{zx} = -1$) would be

$$\begin{aligned} S_{y\beta}^2 - S_{y\alpha}^2 &= -.8S_zS_x + .25(S_{w\beta}^2 - S_{w\alpha}^2), \\ &= S_z(-.8S_x + .5r_{zw\alpha}S_{w\alpha} + .25S_z). \end{aligned}$$

The variance of Y for condition β can clearly be less than at condition α even when $r_{zw\alpha}$ is positive. Generally $r_{zw\alpha}$ would be expected to be negative since $C_{zw\alpha} = C_{w\beta w\alpha} - S_{w\alpha}^2$.

Obviously, many other illustrations could have been given. For example, B_w and/or B_z might have negative signs. The conclusions about the variances would not be changed if there was a mean shift in allocations (i.e., $\bar{Z} \neq 0$). The implications of the model for a given situation can be readily evaluated, however, which makes possible explicit predictions about the effects of any set of Z , on the mean and variance of Y .

In most practical problems more than one measure of different characteristics of the school is apt to be required. If so, the basic model becomes more complicated and terms must be included in the regression equation for each measure. The effect of manipulating any one or combination of these aspects on the mean and variance of the outputs, however, could be estimated in a similar fashion.

REFERENCES

- Blalock, H. M. and Blalock, A. B. *Methodology in social research*, New York: McGraw-Hill, 1968.
- Blalock, H. M. Causal inference, closed population, and measures of association. *American Political Science Review*, 1967, 61, 130-136.
- Cochran, W. G. Errors of measurement in statistics. *Technometrics*, 1968, 10, 637-666.
- Coleman, J. S., et al. Equality of educational opportunity. Washington: U. S. Office of Education, 1966.
- Farrar, D. E. and Glauber, R. R. Multicollinearity in regression analysis: The problem revisited. *Review of Economics and Statistics*, 1967, 49, 92-107.
- Gordon, R. A. Issues in multiple regression. *The American Journal of Sociology*, 1968, 73, 592-616.
- Merton, R. K. In the laboratory world: Reflections on productivity. *Science*, 1968, 160, 639-641.
- Platt, J. R. Strong inference. *Science*, 1964, 146, 347-353.
- Theil, H. Specification errors and the estimation of economic relationships. *Review of the International Statistical Institute*, 1957, 25, 41-51.
- Tukey, J. W. Causation, regression, and path analyses. In Kempthorne, O., et al. (Eds.), *Statistics and Mathematics in Biology* Ames: Iowa State College Press, 1954.

ANALYZING SCHOOL EFFECTS: ANCOVA WITH A FALLIBLE COVARIATE¹

CHARLES E. WERTS AND ROBERT L. LINN

Educational Testing Service

THE analysis of variance, covariance method (ANCOVA) has been employed in nonexperimental school effects studies to control for differential input when studying the differential impact of schools as a categorical treatment factor on some output variable. One of the numerous hazards (Smith, 1957) to interpreting these ANCOVA findings results from the use of input measures known to have considerable errors of measurement. As a consequence, input may not be completely controlled, i.e., the "adjusted" treatment variance which is labeled the "differential" school effect may, to an unknown degree, still reflect differential input. While no general cure for this problem is currently available, it is worth considering what use might be made of reliability estimates and in what circumstance corrections for unreliability would not decrease the estimated differential school effect.

Rationale for the use of ANCOVA

It is, unfortunately, the case that many school effects researchers have used ANCOVA without any substantive justification other than the statement that input should be controlled. It is the case, however, that underlying ANCOVA is a mathematical model, which casts the data in a framework that will be meaningful only to the degree that the phenomena being studied actually behaves like this model. To illustrate this problem consider a case in which

¹ The research reported herein was performed pursuant to Grant No. OEG-1-6-061830-0650 Project No. 6-1830 with the Office of Education, U. S. Department of Health, Education, and Welfare.

there is one input variable X and an output variable Y . The model for ANCOVA is

$$Y_{ij} = A_j + B_w X_{ij} + e_{ij}, \quad (1)$$

where A_j = the Y intercept of the regression line for school j ,

B_w = pooled least squared estimate of the within school regression slope,

and e_{ij} = random fluctuations due to unmeasured factors.

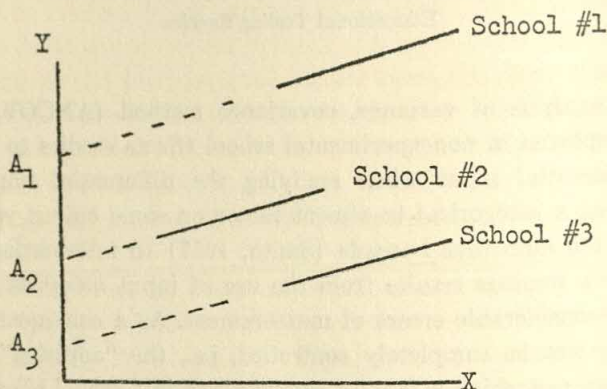


Figure 1. The Model for ANCOVA.

As illustrated in Figure 1, equation (1) when applied to school effects research asserts that the slope of the within school regression lines are all equal. Theoretically, this means that a student entering school #1 will on the average gain $A_1 - A_2$ more units of output than a student entering school #2, regardless of whether his input score is initially high or low. Thus the ordering of the intercepts indicates the ordering (in output units) of the schools in terms of effectiveness, regardless of whether the most effective schools get the students with the highest input scores or not (i.e., whether treatment effects are correlated with the covariate). For any given school the additive constant is the same for everyone; in that sense implying that the school is "equally effective" for students of high as for students of low input scores. If a given school did not actually get students at a particular level of input, it is assumed that if it did, it would do as well with them as with those it

actually received, i.e., that the regression lines may be validly extended beyond their observed ranges.

In actual research the homogeneity of the within group regression slopes can be examined to see if ANCOVA can be applied with this assumption at all. However, the really crucial assumptions will, in general, be untestable, viz., whether the true effect is appropriately simulated by the ANCOVA model, especially whether all other influences operating before, at, or after input are in fact independent of both the treatment effect and the covariates. As a consequence, the interpretation of the ANCOVA results is speculative and should be so labeled. The findings will be useful only to the degree that they serve to engender scientific progress, not as a statement about the existence or nonexistence of school effects. One would be foolish to assert from ANCOVA results that a middle class suburban school would do as well with slum children. Despite these hazards, researchers in the physical sciences have found oversimplified models of reality useful, if only to have something specific to disprove.

For the purpose of considering the implications of random errors of measurement in the covariate we will assume that the ANCOVA model, equation (1), is appropriate, especially the theoretical assumption that a common within group regression slope can be used to adjust for mean differences in the covariate (Smith, 1957). As a consequence the A_j intercepts in equation (1) can be estimated from equation (2):

$$\hat{A}_j = \bar{Y}_j - B_w \bar{X}_j \quad (2)$$

where \bar{Y}_j = sample mean of Y_{ij} for group j ,

and \bar{X}_j = sample mean of X_{ij} for group j .

For heuristic purposes it is useful to substitute equation (2) into equation (1) yielding:

$$Y_{ij} = \bar{Y}_j - B_w \bar{X}_j + B_w X_{ij} + e_{ij} \quad (3)$$

If, in fact, one first computes \bar{Y}_j and \bar{X}_j for each group and assigns these values to individuals in the regression equation

$$Y_{ij} = B_1 \bar{Y}_j + B_2 \bar{X}_j + B_3 X_{ij} + e_{ij},$$

it will be found that the regression coefficients, $B_1 = 1$, $B_2 = -B_3$, and $B_3 = B_w$ (the pooled within group regression slope). This pro-

cedure is a simple transformation of the dummy variable approach to ANCOVA discussed by Cohen (1968) and one could perform similar statistical manipulations. Equation (3) is of heuristic interest because it shows that it is the within groups regression slope, not the between groups or total slope which is the theoretically interesting quantity from the view point of this particular model.

If the most effective schools get the best students (i.e., $r_{AX} > 0$ in equation 1) the between and the total slopes will differ from the within slope (violating the ANCOVA assumption that between and within slopes be homogeneous) but our theory as stated in equation (3) indicates that this will not affect estimates of the intercepts nor, therefore, the ordering of schools for effectiveness. Because of the homogeneity of within group regression, the ordering and variance of the intercepts will be the same as the ordering and variance of the adjusted means. Given that the adjusted means significantly differ from each other, the alternate hypothesis that these differences may be due to unreliability needs to be explored. Because reliability coefficients are usually unavailable, our focus will be on stating the conditions under which it is reasonable to believe that corrections for unreliability should (at least in theory) increase the spread (i.e., the variance) of the adjusted means.

It should be emphasized that equation (2) indicates that the intercepts are a residual quantity representing that part of the output means not accounted for by the input means. Therefore, the finding of a residual school effect (adjusted between school) variance in no sense is positive evidence that the school does have an effect, only that we have not proven that it doesn't. Even if input were adequately controlled, the numerous events happening outside of school during the study might well explain much of the residual variance.

Unreliability

According to the classical theory of unreliability, because of random errors of measurement, a person whose score is deviant on one test will, on the average, have a score closer to the mean on a parallel form of that test (Lord and Novick, 1968, pp. 64-66). One problem in correcting the covariate for unreliability is that a priori we do not know whether on a parallel form the scores for an individual will tend to regress towards the mean of his school (i.e.,

school means are infallible) or towards the overall covariate mean (fallible school means). The former alternative would imply that the school mean itself should be the same on both parallel forms whereas the latter would imply that the school mean itself would "regress" towards the overall mean on the parallel form. In actual school effects studies such alternatives can seldom be tested because of the lack of parallel forms and lack of independence of the measurement errors, from the school effect, other covariates, and the output variable. Furthermore, even if an input variable like social class background could be perfectly measured, this variable is typically used as a surrogate for variables like family values which means that lack of validity will be the more serious hazard to interpretation.

It follows from the above discussion that two cases need to be treated: (a) when the errors of measurement in the covariate are distributed randomly with a zero mean for the total sample irrespective of group, in which case the overall mean on the observed covariate is assumed to be equal to the mean value of the true scores, i.e., the observed group means are assumed to be fallible and (b) when the errors of measurement in the covariate are distributed randomly within groups and have a zero mean within groups, in which case the observed group means are assumed to equal the mean value of the true scores for the persons in that group. For case (a)

- 1.1 The variance of the true covariate scores (σ_x^2) will equal: $\sigma_x^2 = r_{XX}\sigma_X^2$, where r_{XX} = reliability, and σ_X^2 = observed variance of the covariate.
- 1.2 The variance of the true group means ($\sigma_{\bar{x}}^2$) when assigned to individuals in equation (3):

$$\sigma_{\bar{x}}^2 = r_{XX}\sigma_X^2.$$

- 1.3 In terms of true scores indicated by lower case letters,

$$Y_{ij} = B_1\bar{Y}_i + B_2\bar{x}_i + B_3x_{ij} + e_{ij}.$$

- 1.4 The normal equations are:

$$\sigma_{Y\bar{Y}} = B_1\sigma_{\bar{Y}}^2 + B_2\sigma_{\bar{Y}\bar{x}} + B_3\sigma_{\bar{Y}x},$$

$$\sigma_{Y\bar{x}} = B_1\sigma_{\bar{Y}\bar{x}} + B_2\sigma_{\bar{x}}^2 + B_3\sigma_{\bar{x}x},$$

and

$$\sigma_{Yx} = B_1\sigma_{\bar{Y}x} + B_2\sigma_{\bar{x}x} + B_3\sigma_x^2.$$

1.5 To obtain the normal equations in terms of observed scores the following relationships are useful:

- The covariances among the true scores are identical to the covariances among the observed scores (DuBois, 1957).
- When group means are assigned to individuals as in equation (3) the covariance of X with \bar{X} will equal the variance of the group means, i.e., $\sigma_{X\bar{X}} = \sigma_{\bar{X}}^2$ and $\sigma_{Y\bar{Y}} = \sigma_{\bar{Y}}^2$.
- Likewise the covariance of X with \bar{Y} will equal the covariance of \bar{X} and \bar{Y} , i.e., $\sigma_{X\bar{Y}} = \sigma_{\bar{X}\bar{Y}}$.

1.6 By substitution the normal equations become

$$\sigma_{\bar{Y}}^2 = B_1\sigma_{\bar{Y}}^2 + B_2\sigma_{\bar{Y}\bar{X}} + B_3\sigma_{\bar{Y}\bar{X}},$$

$$\sigma_{\bar{Y}\bar{X}} = B_1\sigma_{\bar{Y}\bar{X}} + B_2r_{XX}\sigma_{\bar{X}}^2 + B_3r_{XX}\sigma_{\bar{X}}^2,$$

and

$$\sigma_{YX} = B_1\sigma_{\bar{Y}\bar{X}} + B_2r_{XX}\sigma_{\bar{X}}^2 + B_3r_{XX}\sigma_{\bar{X}}^2.$$

1.7 Solution of these equations yields

$$B_1 = 1, \quad B_2 = -B_3, \quad B_3 = \frac{B_w}{r_{XX}},$$

where B_w = pooled within group regression for observed scores.

1.8 Since the true intercepts equal $B_1\bar{Y}_i + B_2\bar{x}_i$, the variance of the true intercepts (σ_a^2) will be:

$$\sigma_a^2 = B_1^2\sigma_{\bar{Y}}^2 + B_2^2\sigma_{\bar{x}}^2 + 2B_1B_2\sigma_{\bar{Y}\bar{x}}.$$

1.9 By substitution

$$\sigma_a^2 = \sigma_{\bar{Y}}^2 + \frac{B_w^2\sigma_{\bar{x}}^2}{r_{XX}} - 2\frac{B_w}{r_{XX}}\sigma_{\bar{Y}\bar{x}}.$$

1.10 This may be compared with the observed variance among intercepts (σ_A^2),

$$\sigma_A^2 = \sigma_{\bar{Y}}^2 + B_w^2\sigma_{\bar{x}}^2 - 2B_w\sigma_{\bar{Y}\bar{x}}.$$

1.11 Comparison of σ_a^2 to σ_A^2 indicates the "true" variation of the intercepts may be larger or smaller than the observed variation of the intercepts.

1.12 Since $\sigma_{\bar{Y}\bar{x}} \div \sigma_{\bar{x}}^2 = B_{\bar{Y}\bar{x}}$, comparison of step 1.9 to 1.10 indicates that for σ_a^2 to be greater than σ_A^2 either:

- B_w and $B_{\bar{Y}\bar{x}}$ must be opposite sign, or
- if B_w and $B_{\bar{Y}\bar{x}}$ are of the same sign then the absolute value of B_w must be greater than twice the absolute value of $B_{\bar{Y}\bar{x}}$.

When the group mean is the expected mean value of the true scores for the persons in that group:

2.1 The observed variance of the covariate means will equal the variance of the "true" means:

$$\sigma_x^2 = \sigma_{\bar{x}}^2.$$

2.2 When it is assumed that the within group reliability of the covariate scores is the same for all schools, the within group variance (σ_{xw}^2) of the true scores will equal the reliability (r_{xx}) times the within group variance (σ_{xw}^2);

$$\sigma_{xw}^2 = r_{xx}\sigma_{xw}^2 = r_{xx}(\sigma_x^2 - \sigma_{\bar{x}}^2).$$

2.3 It follows that the total variance of the true scores (σ_x^2) is

$$\sigma_x^2 = \sigma_{\bar{x}}^2 + \sigma_{xw}^2 = \sigma_{\bar{x}}^2 + r_{xx}\sigma_{xw}^2.$$

2.4 The regression equation equivalent to equation (3) is (lower cases indicating true scores)

$$Y_{ij} = B_4\bar{Y}_i + B_5\bar{x}_i + B_6x_{ij} + e_{ij}.$$

2.5 The normal equations are:

$$\sigma_{Y\bar{Y}} = B_4\sigma_{\bar{Y}}^2 + B_5\sigma_{\bar{Y}\bar{x}} + B_6\sigma_{\bar{Y}x},$$

$$\sigma_{Y\bar{x}} = B_4\sigma_{\bar{Y}\bar{x}} + B_5\sigma_{\bar{x}}^2 + B_6\sigma_{\bar{x}x},$$

and

$$\sigma_{Yx} = B_4\sigma_{\bar{Y}x} + B_5\sigma_{\bar{x}x} + B_6\sigma_x^2.$$

2.6 By substitution as before

$$\sigma_{\bar{Y}}^2 = B_4\sigma_{\bar{Y}}^2 + B_5\sigma_{\bar{Y}\bar{x}} + B_6\sigma_{\bar{Y}x},$$

$$\sigma_{\bar{Y}\bar{x}} = B_4\sigma_{\bar{Y}\bar{x}} + B_5\sigma_{\bar{x}}^2 + B_6\sigma_{\bar{x}x},$$

and

$$\sigma_{Yx} = B_4\sigma_{\bar{Y}x} + B_5\sigma_{\bar{x}x} + B_6(\sigma_{\bar{x}}^2 + r_{xx}\sigma_{xw}^2).$$

2.7 Solution of these equations yields

$$B_4 = 1, B_5 = -B_6, \text{ and } B_6 = B_w = \text{pooled within slope.}$$

2.8 Since the true intercepts equal $B_4\bar{Y}_i + B_5\bar{x}_i$, the variance of the true intercepts (σ_a^2) will be:

$$\sigma_a^2 = B_4^2\sigma_{\bar{Y}}^2 + B_5^2\sigma_{\bar{x}}^2 + 2B_4B_5\sigma_{\bar{Y}\bar{x}}.$$

2.9 Which by substitution is

$$\sigma_a^2 = \sigma_f^2 + \frac{B_w^2 \sigma_x^2}{r_{xx}} - 2 \frac{B_w}{r_{xx}} \sigma_{fx}.$$

2.10 This, again, may be compared with the observed variance of the intercepts

$$\sigma_A^2 = \sigma_f^2 + B_w^2 \sigma_x^2 - 2B_w \sigma_{fx}.$$

2.11 A comparison of 2.9 and 2.10 indicates that σ_a^2 will be larger than σ_A^2 when

$$\left(\frac{B_w^2 \sigma_x^2}{r_{xx}} - \frac{2B_w}{r_{xx}} \sigma_{fx} \right) > (B_w^2 \sigma_x^2 - 2B_w \sigma_{fx}).$$

2.12 Since $B_{fx} = \sigma_{fx} \div \sigma_x^2$ we find that σ_a^2 is greater than σ_A^2 when

$$\frac{B_w^2}{r_{xx}} - 2 \frac{B_w B_{fx}}{r_{xx}} > B_w^2 - 2B_w B_{fx}.$$

2.13 It can be shown that the inequality in 2.12 will hold if

$$B_w^2(1 + r_{xx}) > 2r_{xx}B_w B_{fx}$$

which for $0 < r_{xx} < 1.0$ is true for the following conditions (a) B_w and B_{fx} have opposite signs or (b) B_w and B_{fx} have the same signs and

$$|B_w| > \frac{2r_{xx}}{1 + r_{xx}} |B_{fx}|.$$

Thus σ_a^2 is always greater than σ_A^2 when B_w and B_{fx} have opposite signs and if B_w and B_{fx} have the same sign then σ_a^2 is greater than σ_A^2 when the absolute value of B_w is greater than the absolute value of B_{fx} multiplied by $2r_{xx}/(1 + r_{xx})$. Since $2r_{xx}/(1 + r_{xx})$ will always be less than 1.0, it follows that where B_w and B_{fx} have the same sign then $\sigma_a^2 > \sigma_A^2$ when $|B_w| > |B_{fx}|$.

Discussion

The above derivations can be summarized as follows:

1. When the internal slope (B_w) and the external slope (B_{fx}) are of opposite sign, then a correction for unreliability should (in theory) result in an increase in the spread of the adjusted means.
2. When the internal slope (B_w) and the external slope are of the same sign, then the effect of correcting for unreliability depends on whether the errors are distributed randomly (and with zero mean)

without respect to groups or are distributed randomly with zero means within groups. In the former case the group means are considered fallible and correcting for unreliability should increase the spread of the adjusted means if the absolute value of the internal slope is at least twice the absolute value of the external slope. In the latter case, the group means are considered infallible and corrections should increase the spread of the adjusted means if the absolute value of the internal slope is at least as large as the absolute value of the external slope multiplied by a fraction which depends on the magnitude of the reliability. Since the former procedure is more conservative than the latter, it should perhaps be preferred in the absence of information about the distribution of the errors. If reliability estimates are available then those values could be substituted into the relevant equations and inferences made about the effect on the spread of the adjusted means. The procedures outlined by Porter (1967) and Thistlethwaite (1968) deal only with the case where the errors of measurement in the covariate are distributed randomly and have a zero mean *within* groups. Our conclusions for this case are in only partial agreement with those of Thistlethwaite who suggested that whenever the absolute value of the internal slope exceeds that of the external slope, reliability corrections will increase the spread of the adjusted means. Our analyses show that the spread of the intercepts will also be increased when the internal and external slopes have opposite signs or when they have the same sign but the absolute value of the internal slopes exceeds the absolute value of the external slope times a fraction that is less than one.

The generalizations in the paper were developed for the case of a single covariate. When multiple covariates are employed a comparison of the magnitude of the variance of the intercepts with and without reliability corrections depends not only on the relative magnitudes of the internal and corresponding external slope but also on the relative magnitudes of the covariances among the covariates. In the case of multiple covariates, it is possible to derive the relevant normal equations as we have done and then to observe the effect of varying reliability values on the spread of the adjusted means.

REFERENCES

- Cohen, J. Multiple regression as a general data-analytic system. *Psychological Bulletin*, 1968, 70, 426-443.

- DuBois, P. H. *Multivariate correlational analysis*. New York: Harper & Bros., 1957.
- Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores*. New York: Addison-Wesley, 1968.
- Porter, A. C. The effects of using fallible variables in the analysis of covariance. (Doctoral dissertation, University of Wisconsin) Ann Arbor, Michigan: University Microfilms, 1967. No. 67-12, 147.
- Smith, H. F. Interpretation of adjusted treatment means and regressions in analysis of covariance. *Biometrics*, 1957, 13, No. 3, 282-308.
- Thistlethwaite, D. L. Analysis of covariance in interpreting group comparisons. Paper presented at the meeting of the American Psychological Association, San Francisco, 1968.

A ONE-WAY ANALYSIS OF VARIANCE FOR SINGLE-SUBJECT DESIGNS

LESTER C. SHINE II AND SAMUEL M. BOWER
The University of Dayton

IN Psychology, there has been a long standing conflict between single-subject and multisubject researchers, which tends to center about the question of whether or not statistics is really useful in single-subject research. Such writers as Sidman (1961) and Skinner (1953) emphasize precisely controlled, single-subject experiments as the most fruitful experimental approach in Psychology, with the role of statistics being limited to the use of elementary descriptive statistics such as means and standard deviations. The powerful, multivariate methods of modern statistics are considered to be inapplicable because they are primarily designed to deal with groups instead of individuals and because their averaging out processes tend to obscure individual differences. Other writers, such as Underwood (1957), argue that the best experimental approach in Psychology is to study groups of subjects to which the modern statistical inference methods may be applied. The purpose of the present paper is to show that it is possible to view certain single-subject designs in such a way as to make the technique of Analysis of Variance applicable to them for the one-way case. Higher ordered designs will be considered in a later paper.

Design and Working Model

The design the authors wish to present for the One-way case is identical in its layout with a standard One-way ANOVA with repeated measures except that instead of using a group of subjects on which is taken a single trial of repeated measures across the levels of the Experimental factor, a single subject is used on which

is taken several trials of repeated measures across the levels of the experimental factor. The two designs are presented schematically in Table 1 for comparison purposes.

In the standard repeated measures design, a pseudo-random factor (Subjects) is usually introduced to allow, with certain assumptions, for the fact that repeated measures are taken on the subjects (Winer, 1962). The purpose of the pseudo-factor is to absorb any correlation between paired columns of measures on subjects, introduced by the presence of effects due to taking repeated measures. It is necessary to assume that the correlation between paired observations under two different levels of the Experimental factor is the same for all possible pairs of such levels, in order for the usual F statistic to be strictly valid (Winer, 1962). Usually, no interaction between the Subject and Experimental factors is permitted

TABLE 1
One Way Repeated Measures Design
B (Conditions)

		1	2	...	q
A (Subjects)	1	X_{11}	X_{12}	...	X_{1q}
	2	X_{21}	X_{22}	...	X_{2q}

	n	X_{n1}	X_{n2}	...	X_{nq}

One Way Single Subject Design
B (Conditions)

		1	2	...	q
A (Trials)	1	X_{11}	X_{12}	...	X_{1q}
	2	X_{21}	X_{22}	...	X_{2q}

	p	X_{p1}	X_{p2}	...	X_{pq}

because such an effect is completely confounded with error effects and cannot, therefore, be estimated separately.

The major assumption that is made for the proposed single subject design is that the subject may be viewed as a response generator, the responses of which to a particular stimulus are statistically independent and normally distributed about a central response value. At first glance this assumption appears to be a contradictory statement since it appears that taking repeated measures on a single subject could very well introduce a correlation between the data under one level of the Experimental factor and the data under another such level. It can be shown, however, that any such correlation can be carried, in a manner similar to that of the standard repeated measures design, by certain effects in the proposed design without affecting the statistical independence of the observations.

Before a detailed examination of the assertions of the preceding paragraph is undertaken, the working model and assumptions for the proposed design will be presented. Under the assumptions that the single subject's responses are statistically independent and normally distributed, the proposed design presented in Table 1 may be represented by the standard model for a fixed factor, two-way ANOVA, with one observation in each cell. The two factors are, as shown in Table 1, the Experimental factor (B) and the Trial factor (A). The mathematical model, definitions, and assumptions are as follows:

$$X_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij}$$

X_{ij} = observation in cell i, j

μ = constant

α_i = constant for each $i = 1, 2, \dots, p$; $\sum_i \alpha_i = 0$

β_j = constant for each $j = 1, 2, \dots, q$; $\sum_j \beta_j = 0$

$(\alpha\beta)_{ij}$ = constant for each i, j ; $\sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$

$\epsilon_{ij} \sim \text{NID}(0, \sigma^2)$ for each i, j

The term α_i is associated with the main effect of the Trial factor, β_j with the main effect of the Experimental factor, $(\alpha\beta)_{ij}$ with the interaction of the two factors, and ϵ_{ij} with all unaccounted or error sources of variation. The term μ is present to allow for the possibility that the average, across cells, of the expected value of a cell observation may not be zero. The use of the preceding model will

certainly raise some serious objections in the reader's mind. The next few paragraphs will attempt to answer these objections.

The first objection that may be raised is that there could be a correlation between the population of the subject's possible responses to a particular treatment on a particular trial and the population of the subject's possible responses to another particular treatment on a particular trial. It is well known that a subject's responses to a stimulus, even the same stimulus under identical conditions, will tend to vary randomly about a central response value, due to such things as random fluctuations in physiological variables over time, random perceptual oscillations over time, etc. If it is assumed that these conditions are operative in single subject research, then the two previously mentioned populations may be assumed to be independent. The random changes would be carried by the ϵ_{ij} term in the model and the central response values would be carried by the other terms in the model.

The second objection that may be raised is that there could be a correlation between the column of data under one treatment and the column of data under another treatment, where the population correlation is defined to be:

$$\rho_{i_1 i_2} = \frac{E \sum_{i=1}^p (X_{i i_1} - \bar{X}_{.i_1})(X_{i i_2} - \bar{X}_{.i_2})}{\sqrt{\left[E \sum_{i=1}^p (X_{i i_1} - \bar{X}_{.i_1})^2 \right] \left[E \sum_{i=1}^p (X_{i i_2} - \bar{X}_{.i_2})^2 \right]}}$$

and the sample correlation is defined to be:

$$r_{i_1 i_2} = \frac{\sum_{i=1}^p (X_{i i_1} - \bar{X}_{.i_1})(X_{i i_2} - \bar{X}_{.i_2})}{\sqrt{\left[\sum_{i=1}^p (X_{i i_1} - \bar{X}_{.i_1})^2 \right] \left[\sum_{i=1}^p (X_{i i_2} - \bar{X}_{.i_2})^2 \right]}}$$

respectively. Under the assumption that the subject's responses are statistically independent, it can be easily shown that:

$$\rho_{i_1 i_2} = \frac{\sum_{i=1}^p [\alpha_i + (\alpha\beta)_{i i_1}][\alpha_i + (\alpha\beta)_{i i_2}]}{\sqrt{\left\{ \sum_{i=1}^p [\alpha_i + (\alpha\beta)_{i i_1}]^2 + (p-1)\sigma^2 \right\} \left\{ \sum_{i=1}^p [\alpha_i + (\alpha\beta)_{i i_2}]^2 + (p-1)\sigma^2 \right\}}}$$

which, as can be seen, may or may not be zero according to the pattern of possible non-error effects. The important point to see

is that $\rho_{i,t,t'}$ can be non-zero, even though the subject's responses are statistically independent, and that its non-zerosness is dependent solely upon, and thus carried by, the pattern of possible effects due to Trials and Trials \times Treatments.

The third objection that may be raised is that there could be a correlation between the row of data for one trial and the row of data for another trial. The situation here is conceptually the same as that of the preceding paragraph with i and j interchanged. Thus, any correlation would be carried by the pattern of possible effects due to Treatments and Trials \times Treatments.

A fourth objection that may be raised is the questioning of the use of a fixed factor for Trials. The trial factor in the proposed design essentially takes the place of the subject factor in a standard, repeated measures design. Since it is usually assumed that the group of subjects in such a design is a random sample, it becomes natural to consider the subject factor to be a random factor. If the trial factor could be considered to be a random factor, then, under the assumptions of this paper, the proposed design could be handled in the same way as a repeated measures design. Thus, the error term for testing the main effect of treatments would simply be the mean square due to the interaction of treatments and trials. The assumptions behind a random factor do not appear, however, to be generally feasible for the proposed design. The inherently sequential nature of the learning process, with which most single subject research is concerned, tends to cause the pattern of central response values under a particular treatment to be non-representative of the pattern that would exist when asymptotic learning has been reached. To meet the conditions for a fixed factor, it is only necessary to assume that the pattern of p central response values, associated with the p trials under a particular treatment, would remain the same if the experiment were theoretically replicated with the same subject under identical experimental conditions. The maintenance of such conditions across replications would, of course, be impossible in the practical sense. The above assumption required for a fixed factor appears reasonable to the present authors and the trial factor may therefore be considered to be fixed. There remains the problem of constructing an appropriate error term for testing the effects in the working model. This problem will be dealt with in the next section.

The Error Term for Testing Effects

The conclusions of the previous section can be summarized by saying that the proposed design may be viewed conceptually as a two-way, fixed factor ANOVA with one observation per cell. Since there is only one observation per cell, the usual within cell estimate of σ^2 cannot be used for testing main effect and interaction sources of variation. It is, therefore, necessary to construct a suitable estimate of σ^2 . If it can be assumed that α_i , the main effect term for trials, changes rather slowly from one trial to the next, then an estimate of σ^2 can be based upon the differences between X_i and X_{i+1} , across odd trials. This assumption seems very reasonable for experiments focusing on learning processes and, of course, could be applicable to many other types of experimental situations.

The estimate of σ^2 that the present authors wish to propose, under the assumption of the preceding paragraph, will be designated MSE' and is defined in the following formula:

$$MSE' \equiv \frac{q \sum_{i=\text{odd}}^{p-1} \frac{1}{2} (\bar{X}_{i+1..} - \bar{X}_{i..})^2}{\frac{p}{2}} = \frac{SSE'}{\frac{p}{2}}, \quad \text{where}$$

$$SSE' \equiv \frac{1}{2q} \sum_{i=\text{odd}}^{p-1} (X_{i+1..} - X_{i..})^2,$$

and where it is understood that an even number of trials, p , has been run. Now, under the assumption that $\alpha_i = \alpha_{i+1}$ for odd i , then $\bar{X}_{i+1..} - \bar{X}_{i..} = \bar{\epsilon}_{i+1..} - \bar{\epsilon}_{i..}$ which is distributed independently, for odd i , and normally with mean zero and variance $(2/q)\sigma^2$. It follows immediately that $E(MSE') = \sigma^2$ and that $(p/2)MSE'/\sigma^2$ is distributed as a chi-square with $p/2$ degrees of freedom. Thus, an approximate F ratio for testing the main effect of conditions would be MSB/MSE' with $(q - 1)$ and $p/2$ degrees of freedom, and an approximate F ratio for testing the interaction of trials and conditions would be $MSAB/MSE'$ with $(p - 1)(q - 1)$ and $p/2$ degrees of freedom. It should be noted that any violation of the assumption that $\alpha_i = \alpha_{i+1}$ for odd i will tend to inflate MSE' and will therefore tend to make the above F ratios conservative.

An appropriate test for testing the main effect of trials is the Mean Square Successive Difference test (Bennet and Franklin, 1961). This test may also be used to test the assumption that

$\alpha_i = \alpha_{i+1}$ for odd i . The formula for the Mean Square Successive Difference (MSSD) Test as applied to the present design is as follows:

$$\eta \equiv \frac{\frac{1}{p-1} \sum_{i=1}^{p-1} (\bar{X}_{i+1..} - \bar{X}_{i..})^2}{\frac{1}{p-1} \sum_{i=1}^p (\bar{X}_{i..} - \bar{X}_{..})^2} = \frac{\text{MSSD}}{\text{MSA}},$$

where

$$\text{MSSD} \equiv \left[(1/q) \sum_{i=1}^{p-1} (X_{i+1..} - X_{i..})^2 \right] / (p-1)$$

and MSA is the usual mean square associated with the main effect of trials. Critical values for η for 5 per cent and 1 per cent one tailed tests have been summarized by Bennett and Franklin (1961). For $p > 25$ a z -test may be used by employing the following formula:

$$z = (1 - \eta/2) \sqrt{(p-1)(p+1)/(p-2)}.$$

A significant η or z in either direction is evidence for a significant main effect for trials. A *right tailed* significant η or a *left tailed* significant z is evidence for rejecting the assumption that $\alpha_i = \alpha_{i+1}$ for odd i .

Schematic Calculation Procedures

As before, let A stand for the trial pseudofactor and B stand for the experimental factor. A schematic ANOVA Source Table for the proposed design is presented in Table 2.

First, lines A, B, AB, and Total may be filled out, except for the F and η columns, by performing a standard, Two-way Fixed Factor ANOVA, with one observation per cell, on the data as if there were no repeated measures. Second, lines SD and E' may be filled out by applying the formulas given at the base of the Source Table presented in Table 2. Third, the slot for η and the slots for the two F ratios may be filled out according to the formulas given in the body of the Source Table. Fourth, A is considered to be a significant source of variation if η (or z) is significant. Fifth, if η is not significant in the right tail (or z is not significant in the left tail), then the F ratios for the B and AB sources of variation may be considered reasonably valid and may therefore be interpreted

TABLE 2
ANOVA Source Table ($p = \text{even only}$)

Source	SS	df	MS	F	η^a
A	SSA	$p - 1$	MSA	—	MSSD/MSA
	SD	—	MSSD	—	—
B	SSB	$q - 1$	MSB	MSB/MSE'	—
AB	SSAB	$(p - 1)(q - 1)$	MSAB	MSAB/MSE'	—
	E'	$p/2$	MSE'	—	—
Total	TSS ^b	$pq - 1^b$	—	—	—

$$\text{MSSD} = \frac{\frac{1}{q} \sum_{i=1}^{p-1} (X_{i+1..} - X_{i.})^2}{p - 1} = \frac{1}{q(p - 1)} [(X_{2.} - X_{1.})^2 + (X_{3.} - X_{2.})^2 + \cdots + (X_{p.} - X_{p-1..})^2]$$

$$\text{SSE}' = \frac{1}{2q} \sum_{i=\text{odd}}^{p-1} (X_{i+1..} - X_{i.})^2 = \frac{1}{2q} [(X_{2.} - X_{1.})^2 + (X_{4.} - X_{3.})^2 + \cdots + (X_{p.} - X_{p-1..})^2]$$

^a May be replaced by $z = (1 - \eta/2) \sqrt{(p - 1)(p + 1)/(p - 2)}$, if $p > 25$.

^b The total does not include the figures for E'.

in the usual manner. If η is significant in the right tail (or z in the left tail), then there is evidence that the assumption $\alpha_i = \alpha_{i+1}$ for odd i is invalid, and the F ratios should, therefore, be considered to be invalid. Since, in this case, the F ratios will tend to be conservative, the individual experimenter may, based upon his individual situation, decide to accept the F ratios anyway.

Summary

The purpose of the present paper is to show that a one-way ANOVA may, under certain assumptions, be applied to a single-subject design in which one subject is observed on several trials under each of several experimental conditions. It is assumed that the subject may be viewed as a response generator the responses of which to a particular stimulus are statistically independent and normally distributed about a central response value. A fixed, trial pseudo-factor is introduced to carry any correlations introduced

in the data by using only one subject. It is shown that the one-way design may be handled as a two-way design, with one observation per cell, for which, under the assumption that the main effects for trials change slowly from one trial to the next, a modified error term for testing effects may be constructed. A statistical test is given for testing the preceding assumption, and schematic calculation procedures are presented.

REFERENCES

- Bennett, C. A. and Franklin, N. L. *Statistical analysis in chemistry and the chemical industry*. New York: John Wiley and Sons, Inc., 1961.
- Sidman, M. *Tactics of scientific research*. New York: Basic Books, Inc., 1961.
- Skinner, B. F. *Science and human behavior*. New York: McMillan, 1953.
- Underwood, B. J. *Psychological research*. New York: Appleton-Century Crofts, Inc., 1957.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

AN EMPIRICAL NOTE ON CORRELATION COEFFICIENTS CORRECTED FOR RESTRICTION IN RANGE

ROBERT A. FORSYTH
University of Iowa

BEHAVIORAL scientists frequently utilize interval estimation techniques and point estimates of parameters for which there exists only relatively inadequate sampling theories. Correlation coefficients corrected for attenuation, for example, have estimated standard error formulas, but the form of the sampling distribution is not known. Investigators using such coefficients often employ the sampling theory for the Pearson product-moment correlation coefficient (Yamamoto, 1965). The inadequacies associated with this procedure have been discussed recently by Forsyth and Feldt (1969).

A fairly similar situation prevails when an experimenter wishes to estimate the correlation between variables X and Y for some population but complete data are available for only a select, non-random sample of subjects. In a common example, subjects have been selected on the basis of their X -scores and data for the Y -variable are available only for this restricted group. This situation would occur if a college entrance test was utilized as a selection device and it was desired to estimate the correlation between the test and freshman grade point average for a population of students represented by all those who applied for admission. In this case, data on the test are available for both groups, but grade point data are available for the restricted group only. A formula is available which, under certain assumptions, yields an estimate of the correlation for the complete population. This is sometimes labeled the correlation coefficient corrected for restriction in range. (Actually there are several slightly different situations to which the restriction

in range label applies. However, the situation described above is the most common.) Both Gulliksen (1950, pp. 136-137) and Lord and Novick (1967, pp. 142-143) have derived the following formula for estimating the unrestricted correlation between X and Y when the variance of X is known for both the restricted and unrestricted groups:

$$r_{xy} = \frac{s_x r_{x^*y^*}}{\sqrt{s_x^2 r_{x^*y^*}^2 - s_{x^*}^2 - s_x^2 r_{x^*y^*}^2}} \quad (1)$$

where

s_x^2 = variance of the unrestricted group.

$s_{x^*}^2$ = variance of the restricted group.

$r_{x^*y^*}$ = correlation between x and y for the restricted group.

r_{xy} = correlation between x and y for the unrestricted group.

In addition to the assumptions of linearity and homoscedasticity the population counterparts of the sample estimates must be employed in the derivation of this equation. If the assumptions are met and if population facts are known, then Formula (1) will give the exact value of ρ_{xy} , i.e., the population correlation for the unrestricted population. However, population facts are never known and hence sample estimates must be utilized in the formula. Consequently, this raises the issue of sampling error in the estimate, an issue which can be resolved only by a consideration of the sampling distribution for the r_{xy} obtained from Equation (1). The form of this sampling distribution is not known nor are any standard error formulas available.

Although many investigators may merely desire point estimates of these correlation coefficients corrected for restriction in range (Humphreys, 1968) some researchers (Yamamoto, 1965) may wish to establish confidence intervals or test hypotheses about the population value of the corrected r .

The primary purpose of this study was to examine empirical sampling distributions of corrected r 's obtained via Equation (1). The investigation was not concerned with the appropriateness of Equation (1) for yielding "good" point estimates of ρ_{xy} . Rather, it was concerned with the accuracy of the nominal confidence coefficient of the obtained confidence intervals for ρ_{xy} when Equation (1) was utilized and when it was assumed that these corrected r 's were distributed as Pearson product moment r 's. When it was

found that the traditional theory was grossly inadequate, an attempt was made to develop a more appropriate technique.

Procedures

Corrected r 's were considered for three sample sizes for the restricted group (25, 50, and 100), four cutoff points (P_{10} , P_{25} , P_{50} , and P_{75}), and two population values of the correlation (.80 and .50). Thus, 24 different empirical sampling distributions were generated utilizing the procedure described below.

A computer program was written to yield normally distributed scores X and Y with zero mean and unit variance and predetermined linear correlation ρ_{xy} . For each pair of X and Y values generated, the X -value was compared to a given cutoff point. If the X -score was greater than the cutoff point, the pair of scores was utilized in the analysis. This process was continued until the desired sample size was attained. Once the sample was completed, the necessary statistics for the right side of Equation (1) were calculated and then the equation was solved for r_{xy} . This process was repeated 1000 times to form an empirical sampling distribution of corrected r 's with a specified N for the restricted group, a specified cutoff point, and a specified population r .

It should be noted that this procedure made the cutoff point a fixed parameter but allowed the sample size of the unrestricted group to vary. This meant that the variance estimate of X for the unrestricted group (s_x^2) was based on different sample sizes from one replication to another. Of course, the sample size of the restricted group was constant over all replications. (At this time it should be indicated that several empirical distributions of corrected r 's were obtained using the population value of σ_x^2 in Equation (1). The use of σ_x^2 in this equation would be feasible, for example, in the entrance test situation described above. In that instance the investigator could use the test norms, under certain assumptions, to obtain σ_x^2 instead of obtaining s_x^2 from the unrestricted sample. This, of course, solves the problem associated with the computing of s_x^2 from different size samples in the same sampling distribution. However, the results of these distributions were very similar to the corresponding distributions using s_x^2 .)

After the empirical sampling distributions were obtained, the means and variances were computed for each of the 24 distri-

butions. Finally, Fisher's z_r -transformation was utilized to establish three confidence intervals (.90, .95, .99) for ρ_{xy} for each of the 1000 statistics in a given empirical sampling distribution. The proportion of these intervals which contained ρ_{xy} was calculated and compared to the nominal γ -level. The difference between these two values was used as an index of the amount of error present when the traditional Pearson product moment confidence interval procedure is utilized with restricted r 's. These differences were obtained for all 24 empirical distributions.

Results and Discussion

The means and variances of the 24 empirical sampling distributions along with the means and variances of corresponding Pearson product moment sampling distributions are given in Table 1.

It can easily be seen from Table 1 that there is a tendency for the means of the empirical sampling distributions to be less than the corresponding means of the Pearson product moment distribution. However, except for two means (.468 and .444) when $\rho_{xy} = .50$ and $N = 25$ the discrepancy was not too great. As the sample size increased, the means of the empirical distributions became very similar to the product moment mean. However, it is equally apparent from Table 1 that the variances of the empirical distributions differed markedly from the variances of the corresponding product moment distributions. Furthermore, these discrepancies did not decrease as sample size increased. For example, when $N = 25$ and $\rho_{xy} = .50$ the ratio of the largest corrected r variance and the Pearson r variance was 3.83. When $N = 50$ and $\rho_{xy} = .50$ this same ratio was 3.52 and when $N = 100$ and $\rho_{xy} = .50$ the ratio was 3.62. The corresponding ratios when $\rho_{xy} = .80$ were 3.01, 2.29, and 2.21, respectively.

There is some indication that as ρ_{xy} increases, the various cutoff points have less influence of the variability of the sample estimate. However, it is obvious that the variances of the corrected r 's become increasingly greater as the cutoff point becomes increasingly higher regardless of the magnitude of ρ_{xy} .

In general, the results given in Table 1 support the contention that the utilization of Fisher's z -transformation to establish confidence intervals for ρ_{xy} will prove to be relatively inaccurate. The extent

TABLE 1
Means and Variances for Sampling Distributions

<i>N</i>	ρ_{xy}	Cutoff	Mean	Variance
25	.50	—	.492	.0247
25	.50	P_{15}	.484	.0331
25	.50	P_{25}	.491	.0210
25	.50	P_{35}	.498	.0534
25	.50	P_{45}	.444	.0945
25	.80	—	.794	.0363
25	.80	P_{15}	.794	.0371
25	.80	P_{25}	.789	.0393
25	.80	P_{35}	.789	.0128
25	.80	P_{45}	.784	.0190
50	.50	—	.496	.0118
50	.50	P_{15}	.493	.0149
50	.50	P_{25}	.486	.0198
50	.50	P_{35}	.483	.0287
50	.50	P_{45}	.488	.0415
50	.80	—	.797	.0028
50	.80	P_{15}	.796	.0033
50	.80	P_{25}	.793	.0041
50	.80	P_{35}	.796	.0044
50	.80	P_{45}	.795	.0064
100	.50	—	.498	.0058
100	.50	P_{15}	.496	.0069
100	.50	P_{25}	.491	.0095
100	.50	P_{35}	.497	.0131
100	.50	P_{45}	.491	.0210
100	.80	—	.799	.0014
100	.80	P_{15}	.799	.0015
100	.80	P_{25}	.796	.0019
100	.80	P_{35}	.797	.0023
100	.80	P_{45}	.798	.0031

* Values for the mean and variance of these Pearson product moment sampling distributions are taken from Soper et al., (1916, pp. 371-372).

of this inaccuracy was determined by the following technique. In each empirical sampling distribution of corrected r 's, all r -values were converted to z_r -values via the formula, $z_r = .5 \log_e [(1+r)/(1-r)]$. Confidence intervals for z_r were obtained by utilising the traditional standard error of these z_r -values, namely $1/\sqrt{N-3}$. The proportion of these 1000 intervals which enclosed z_r was calculated. (Of course, this is the same proportion of confidence intervals which would enclose ρ_{xy} if the limits of the interval were converted to correlation

values.) These results are shown as part of Table 2 (in the column under traditional standard error).

A comparison of the empirical proportion with the nominal γ -level provides a basis for determining the accuracy of the traditional technique. The data in Table 2 furnish very strong evidence that fairly gross errors will be made if the traditional Pearson r sampling theory is employed for estimated r 's. When the nominal $\gamma = .90$, the empirical estimates of γ ranged from .616 to .895. When $\gamma = .95$, the range was from .703 to .942. Finally, when $\gamma = .99$, the estimated γ 's ranged from .818 to .989. Over all 24 distributions the empirical γ 's were on the average .11 units below .90, .09 units below .95 and .05 units below .99.

TABLE 2
Proportion of Confidence Intervals Enclosing ρ_{ZY}

N	ρ_{ZY}	Cut-off	Nominal Confidence Interval					
			$\gamma = .90$		$\gamma = .95$		$\gamma = .99$	
			Traditional Standard Error	Adjusted Standard Error	Traditional Standard Error	Adjusted Standard Error	Traditional Standard Error	Adjusted Standard Error
25	.50	P_{10}	.852	(.897) ^a	.919	(.936) ^a	.969	(.980) ^a
50	.50	P_{10}	.854	(.891)	.920	(.941)	.979	(.985)
100	.50	P_{10}	.868	(.903)	.925	(.940)	.973	(.982)
25	.80	P_{10}	.877	(.913)	.933	(.955)	.985	(.993)
50	.80	P_{10}	.895	(.919)	.942	(.962)	.980	(.992)
100	.80	P_{10}	.883	(.912)	.936	(.961)	.989	(.997)
25	.50	P_{15}	.837	(.892) ^b	.888	(.953) ^b	.961	(.989) ^b
50	.50	P_{15}	.834	(.873)	.866	(.936)	.953	(.985)
100	.50	P_{15}	.790	(.875)	.870	(.935)	.958	(.982)
25	.80	P_{15}	.859	(.920)	.910	(.952)	.968	(.993)
50	.80	P_{15}	.845	(.918)	.908	(.960)	.976	(.994)
100	.80	P_{15}	.858	(.924)	.920	(.958)	.969	(.992)
25	.50	P_{20}	.714	(.873) ^c	.801	(.930) ^c	.899	(.977) ^c
50	.50	P_{20}	.714	(.871)	.798	(.925)	.911	(.978)
100	.50	P_{20}	.741	(.869)	.813	(.920)	.907	(.978)
25	.80	P_{20}	.790	(.944)	.883	(.973)	.959	(.993)
50	.80	P_{20}	.818	(.940)	.890	(.977)	.971	(.995)
100	.80	P_{20}	.798	(.914)	.878	(.967)	.955	(.994)
25	.50	P_{75}	.616	(.872) ^d	.703	(.929) ^d	.818	(.984) ^d
50	.50	P_{75}	.652	(.859)	.716	(.920)	.836	(.977)
100	.50	P_{75}	.623	(.846)	.716	(.910)	.835	(.977)
25	.80	P_{75}	.751	(.957)	.823	(.978)	.924	(.996)
50	.80	P_{75}	.751	(.947)	.837	(.969)	.936	(.995)
100	.80	P_{75}	.735	(.931)	.809	(.970)	.919	(.993)

^a Standard error for z_r when P_{10} is the cutoff point is: $1/\sqrt{N} - .15N - 3$.

^b Standard error for z_r when P_{15} is the cutoff point is: $1/\sqrt{N} - .30N - 3$.

^c Standard error for z_r when P_{20} is the cutoff point is: $1/\sqrt{N} - .45N - 3$.

^d Standard error for z_r when P_{75} is the cutoff point is: $1/\sqrt{N} - .60N - 3$.

Three definite trends can be seen for these data (traditional standard error data): (1) for a given ρ_{xy} and a given cutoff point the amount of error is relatively independent of sample size; (2) as the selection procedure becomes more stringent the amount of error increases; and (3) as ρ_{xy} increases the magnitude of the errors decreases.

Approximate Confidence Interval Techniques

When the empirical sampling distributions of the corrected r 's were obtained, they were compared to the corresponding Pearson product moment r distribution (i.e., a distribution with the same ρ_{xy} and the same sample size). It was very apparent, as Table 1 would indicate, that the corresponding pairs of distributions were markedly different. However, when the corrected r distributions were compared to Pearson r distributions for the same ρ_{xy} value but for a smaller sample size the two distributions become fairly similar. For example, the corrected r distribution for $\rho_{xy} = .50$, $N = 25$, and cutoff = P_{50} is more similar to a Pearson r distribution for $\rho_{xy} = .50$ and $N = 15$ than it is to the sampling distribution for $N = 25$. These facts suggested that it might be possible to find reasonable confidence intervals by adjusting the N size in the traditional standard error formula for z_r . Of course, as Table 2 indicates, such adjustment would be directly related to the degree of restriction. That is, the higher the restriction the greater the decrease in sample size. However, it is also evident in Table 2 that such a technique would be somewhat inadequate since the accuracy of the traditional standard error formula for z_r is related to the value of ρ_{xy} . Since merely adjusting the sample size in the traditional formula does not provide for varying ρ_{xy} values, any techniques using this simple approach would not be completely adequate. Nevertheless, it was decided to investigate the effectiveness of adjusting the traditional standard error formula for z_r .

A trial and error procedure was initiated to find these adjusted standard errors. The adjusted values which seemed to provide reasonable confidence intervals are given in the footnote to Table 2. The actual proportion of confidence intervals which enclosed ρ_{xy} when these adjusted values were used are shown in Table 2. There is little doubt that the adjusted standard errors for z_r provided much more accurate confidence intervals when compared to the traditional method. It must be emphasized, however, that these

standard error formulas are merely empirical values which seem to hold for the 24 distributions under study.

An attempt was made to see how well these standard errors would function in three new sampling distributions (1000 estimates in each). These three distributions are identified in Table 3. It may be noted that entirely different values of N , ρ_{xy} , and ρ_z were utilized. The adjusted standard error formulas for these distributions were obtained by a linear interpolation procedure utilizing the adjusted standard errors in Table 2. An example will illustrate the procedure. The adjusted standard error for z_r when P_{35} is the cutoff point is $1/\sqrt{N - .30N - 3}$. The adjusted standard error when P_{50} is the cutoff point is $1/\sqrt{N - .45N - 3}$. Therefore, when P_{40} is the cutoff point the correction factor is found by solving the following expression, $3/5(.45N - .30N) + .30N$. Thus, when P_{40} is the cutoff the adjusted standard error is $1/\sqrt{N - .39N - 3}$. The adjusted standard error when P_{50} was utilized as the cutoff was found by assuming that the denominator of the standard error formulas must equal zero when no one is selected (i.e., P_{100} is cutoff point).

After the adjusted standard errors were computed, 1000 confidence intervals for z_r were computed in each of the three distributions. The proportion of these intervals which would enclose ρ_{xy} were found. These proportions are shown in Table 3. The proportion of confidence intervals which would enclose ρ_{xy} using $1/\sqrt{N - 3}$ as the standard error are also presented.

TABLE 3

Cross Validation Distributions: Proportion of Confidence Intervals Enclosing Population Correlation Coefficient

N	ρ_{xy}	Cut-off	Nominal Confidence Interval				
			$\gamma = .90$		$\gamma = .95$		$\gamma = .99$
			Adjusted Standard Error	Traditional Standard Error	Adjusted Standard Error	Traditional Standard Error	Adjusted Standard Error
40	.60	P_{40}	.877 ^a	(.776)	.934	(.842)	.992
75	.70	P_{50}	.893 ^b	(.726)	.944	(.817)	.985
40	.65	P_{50}	.964 ^c	(.590)	.978	(.665)	.996

^a Estimated standard error of $z_r = 1/\sqrt{N - .39N - 3}$.

^b Estimated standard error of $z_r = 1/\sqrt{N - .51N - 3}$.

^c Estimated standard error of $z_r = 1/\sqrt{N - .79N - 3}$.

The results shown in Table 3 are very encouraging. The adjusted standard error technique seems to give reasonably accurate confidence intervals. Therefore, it is recommended, (assuming that the bivariate distribution does not deviate markedly from bivariate normal) that if an investigator wishes to establish a confidence interval for ρ , based on an estimate provided by Equation (1), then the approximate technique described in this paper should be utilized until a more accurate sampling theory is developed.

REFERENCES

- Forsyth, Robert A. and Feldt, Leonard S. An investigation of empirical sampling distributions of correlation coefficients corrected for attenuation. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 61-72.
- Gulliksen, Harold. *Theory of mental tests*. New York: John Wiley and Sons, Inc. 1950.
- Humphreys, Lloyd G. The fleeting nature of the prediction of college and academic success. *Journal of Educational Psychology*, 1968, 59, 375-380.
- Lord, Frederic M. and Novick, Melvin R. *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley Publishing Company, 1968.
- Soper, H. E., Young, A., Cave, B. Lee, A. and Pearson, K. On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "student" and R. A. Fisher. *Biometrika*, 1916, 11, 328-413.
- Yamamoto, Karou. Effects of restriction of range and test unreliability on correlation between measures of intelligence and creative thinking. *British Journal of Educational Psychology*, 1965, 35, 300-305.

A COMPARISON OF COMPUTER-SIMULATED CONVENTIONAL AND BRANCHING TESTS¹

CARRIE WHERRY WATERS

Center for Psychological Services
Ohio University, Athens

A. G. BAYROFF

U. S. Army Behavioral Science
Research Laboratory
Arlington, Virginia

In the usual testing situation, each examinee takes all of the items and item sequence is the same for each examinee. It is possible, however, to have sequential or branching tests in which all examinees do not take the same items and the sequence of item presentation for an individual is some function of his performance on previous items. The rationale for this latter procedure is that the presentation of items based on an examinee's past performance allows each individual to take items which are progressively more appropriate to his own level of ability. It is conceivable that such a procedure would reduce testing time and for a given amount of time would permit more accurate measurement of an individual's ability.

Krathwohl and Huyser (1956) reported the development of a 6-item per subject branching test for college students which correlated .78 with a 60-item parent test and .68 with a reading test. Two experimental 6-item per subject branching tests for Army enlisted personnel (Bayroff, Thomas, and Anderson, 1960; Seeley, Morton, and Anderson, 1962) each correlated .63 with a 25-item conventional test measuring similar content. Assuming item inter-

¹ The opinions expressed in this paper are those of the authors and do not necessarily reflect official Department of the Army policy.

correlations of .64, Waters (1964) compared a hypothetical 5-item per subject branching test with four hypothetical 5-item conventional tests, which differed in item difficulty distributions, and found the branching test correlated slightly higher with an underlying ability criterion than did any of the conventional tests regardless of whether free response or multiple choice format was assumed.

Tests

Conventional Tests

Five-, ten-, and fifteen-item hypothetical conventional (C) tests were evaluated. All tests were symmetric around $p = .50$, but varied in item difficulty distributions. The distributions investigated were all items at $p = .50$ (C50), roughly normal (CN), or rectilinear (CR). Each of the CN and CR tests was tried out with difficulty ranges of .30 through .70 and .10 through .90. Table 1 gives the C50, CN, and CR item difficulty distributions for the five-, ten-, and fifteen-item conventional tests.

Branching Tests

One-Item Per Stage: Six hypothetical 1-item per stage (1-PS) branching tests were evaluated. Two tests were studied at each of the three test lengths (5, 10, and 15 items). One of the two tests covered a difficulty range of .30 through .70 and the other ranged

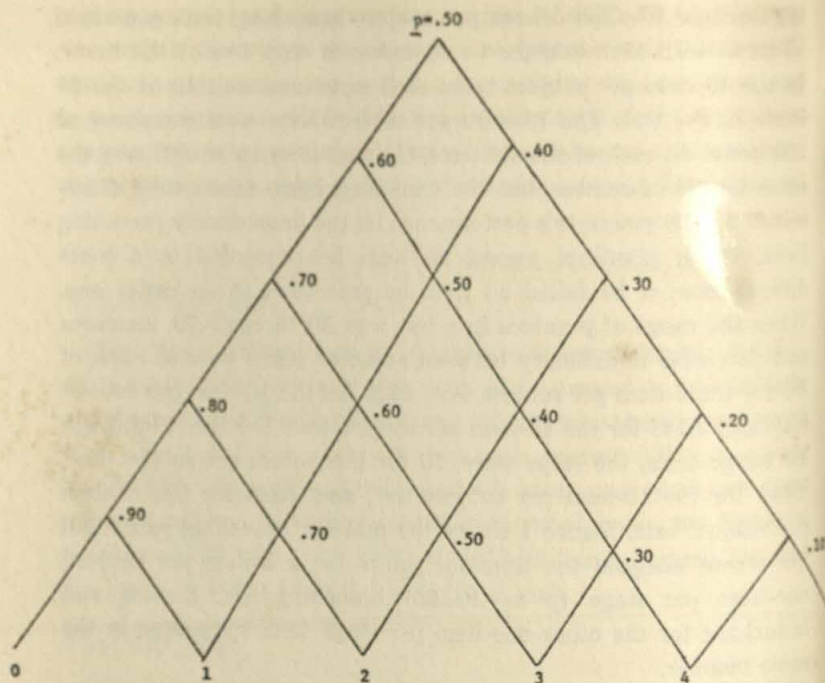
TABLE 1
Number of Items at Each p -Value for Conventional Tests

p -Value	5-Item Tests					10-Item Tests					15-Item Tests				
	all .50	.30-.70 Range		.10-.90 Range		all .50	.30-.70 Range		.10-.90 Range		all .50	.30-.70 Range		.10-.90 Range	
		N	R	N	R		N	R	N	R		N	R	N	R
.90				—	1				1	2				2	
.80															
.70		—	1	—	1		1	2	2	2		2	3	3	
.60		—	1				2	2				3	3		
.50	5	—	1	—	1	10	4	2	4	2	15	5	3	5	
.40		—	1				2	2				3	3		
.30		—	1	—	1		1	2	2	2		2	3	3	
.20															
.10				—	1				1	2				2	

* No 5-item normally distributed tests were evaluated.

.10 through .90. The 5-item per subject branching tests contained 15 items with each examinee responding to only five of the items. In the 10-item per subject tests, each examinee took 10 of the 55 items in the test. The 15-item per subject tests were composed of 120 items. In each of the six tests, the first item ($p = .50$) was the same for all examinees, but the remaining items taken were determined by the examinee's performance on the immediately preceding item. If an examinee passed an item he proceeded to a more difficult one; if he failed an item he proceeded to an easier one. When the range of p -values in a test was .30 through .70, increases and decreases in difficulty between adjacent items were in steps of .05 for the 5-item per subject test, .0222 for the 10-item per subject test, and .0143 for the 15-item per subject test. For the .10 through .90 range tests, the steps were .10 for the 5-item per subject test, .0444 for the 10-item per subject test, and .0286 for the 15-item per subject test. Figure 1 shows the possible branching paths and the scores assigned the terminal points for a 5-item per subject, one-item per stage ($p = .10-.90$) branching test. Scoring and branching for the other one-item per stage tests were done in the same manner.

Two-Item Per Stage: Four hypothetical 2-item per stage (2-PS), 10-item per subject branching tests were evaluated. Each of these tests was composed of 114 items. At each stage in these tests the examinees took two items of the same difficulty level. The first two items taken by all examinees had p -values of .50. If the examinee passed both items in a pair he branched to a more difficult item pair; if he passed one of the items in a pair he branched to a pair of equal difficulty; if he failed both items in a pair he proceeded to an easier pair of items. Items for two of the tests covered a difficulty range of .30 through .70, while the other two tests ranged .10 through .90. For each of these difficulty ranges, one branching test was developed by having equally spaced item pairs in the terminal row of the test (2-PS-E). The p -values of the item pairs in the other rows were determined from the terminal item pair values. For the other 2-item per stage tests, 2-PS-U, (one for each of the item difficulty ranges) the item pair p -values were determined by branching downward from the $p = .50$ item pair to the terminal row of item pairs. Using this procedure, the item pairs in the terminal rows were not equally spaced as in the 2-PS-E tests but



Scoring Scale

Figure 1. Five-item per subject, one-item per stage branching test.

were spaced so that the intervals between item pairs were smaller in the middle part of the difficulty ranges, and larger nearer the extreme difficulty values. Scores for all four of the two-item per stage tests ranged from 0 to 62.

Computational Procedures and Assumptions

Statistical computations were based on a theoretical model presented by Lord (1952). The model assumes that there is a trait or ability underlying the raw scores on a test, and that the probability of an examinee's responding correctly to a test item is a normal ogive function of his position on the ability dimension. Since item responses are a function only of scores on the ability continuum, they are independent of each other when ability is held constant. When all of the items in a test are assumed to have the same

biserial (R_i) with ability, R_i^2 is an estimate of item intercorrelation. Three major steps are involved in obtaining the correlation between test score and underlying ability: the proportion of examinees passing each item is determined for each of the ability levels under consideration; the conditional distribution of test scores is obtained for each ability level; and the bivariate frequency distribution of test score and ability is obtained.

*Proportion of Examinees at a Given level of Ability
Who Pass an Item*

When the group tested is assumed to be normally distributed on ability, Lord's formulas (9) and (10) may be used to find the proportion of examinees who pass each of the test items when ability is held constant. In Lord's notation, a value of g_i (the z score corresponding to the P -value of item i at a specified ability level) is computed for each ability level under consideration by formula (9):

$$g_i = \frac{h_i - R_i \cdot c}{K_i} \text{ where,}$$

h_i = the z value corresponding to the population p -value of item i

R_i = the biserial correlation between item i and underlying ability

c = the z score representing the ability level being considered

$$K_i = \sqrt{1 - R_i^2}$$

Each g_i is converted to P_i (P -value of item i for examinees at a given ability level) by Lord's formula (10):

$P_i = A(g_i)$ = area of normal curve above the point g_i . These P_i values are computed for ability levels.

*Conditional Test Score Distribution for Given
Ability Levels*

For conventional tests, the distribution of test scores at each of the specified ability levels may be computed by expansion of Lord's formula (11):

$$\prod_{i=1}^n (P_i + Q_i) \text{ where,}$$

$\prod_{i=1}^n$ indicates the successive multiplication of the $(P_i + Q_i)$ terms
 n = number of items in test

P_i = proportion passing item i for the given ability level

$Q_i = 1 - P_i$

Terms of this expansion give all possible ways of obtaining various test scores. Those terms which lead to the same test score are summed to obtain the distribution of test scores for a given ability level.

Although Lord does not discuss branching tests, his model is also applicable to this type of test. For a branching test the proportion of examinees (at a specified ability level) following any path may be determined by multiplying the P_i or Q_i values (as obtained by Lord's formulas 9 and 10) of the items which make up that path. If an item is passed its P_i value is used, if an item is failed its Q_i value is used. Such a proportion is computed for each path and values for paths leading to the same test score are summed to obtain the test score distribution.

Bivariate Frequency Distribution of Test Score and Ability

For both conventional and branching tests, the bivariate distribution of test score and ability is obtained by multiplying the conditional test score distribution for each ability level by the ordinate value of the normal curve at that ability level (Lord's formula 14, applicable when a normal distribution of ability is assumed). The test-ability correlation coefficient may be computed from this scatterplot.

A FORTRAN program which performs these computations was written for the GE 225 computer by Mr. Sidney Sachs of the Computer Applications Branch, U. S. Army Behavioral Science Research Laboratory. This program was used to obtain the test-ability coefficients reported in this study. It should be noted that Brogden (1946), Tucker (1946), and Lord (1952) have provided computationally easier formulas for obtaining the test-ability coefficients for conventional tests.

In this study, the distribution of underlying ability in the theoretical sample of examinees was assumed to be normal with $\bar{X} = 0$ and $\sigma = 1.00$. Twenty-nine levels of ability, measured in standard scores ranging from +3.5 to -3.5 in steps of .25 were used. The biserial correlation between an item and ability was constant for all items in a given test. For each of the 5- and 10-item per subject

conventional and branching tests evaluated, the assumed biserial was varied from .30 to .90 in steps of .10. The 15-item tests were evaluated at biserials of .40, .60 and .80.

Results and Discussion

Five-Item Tests

Conventional Tests. The correlation coefficients between test score and ability for the 5-item conventional tests are shown in the first three rows of Table 2. For biserials of .30 through .70 ($r_{ij} = .09$ through .49), the all .50 (C50) test obtained the highest coefficients, and the .3 through .7 CR test yielded a higher relationship to the ability criterion than did the .1 through .9 CR test. At the .80 biserial ($r_{ij} = .64$), the .3 through .7 CR test yielded the highest coefficient, and the C50 test was next. Finally, at the assumed biserial of .90 ($r_{ij} = .81$) the wide range rectilinear test (.1 through .9 CR test) had the highest correlation coefficient, and the C50 test had the lowest coefficient of the three conventional tests. Overall, the C50 was best for low to moderate item intre correlations; the moderate range (.3 through .7) and eventually the wider range (.1 through .9) tests were best for higher intercorrelations.

TABLE 2
*Test Score-Ability Correlation Coefficients for 5-Item Per Subject
Conventional and Branching Tests*

Biserial	.30	.40	.50	.60	.70	.80	.90
Test							
C (all .50)	482*	601	696	769	823	858	871
C (.3-.7, R)	473	591	686	762	819	861	887
C (.1-.9, R)	434	549	646	726	793	850	900
B (.3-.7, 1-PS)	478	599	694	774	835	880	906
B (.1-.9, 1-PS)	461	580	680	760	826	878	920

* Decimal points omitted.

Branching Tests. The results for the 5-item per subject branching tests are shown in the last two rows of Table 2. The coefficient for the moderate range .3 through .7 test was higher than that for the wider range test for assumed biserials of .30 through .80; the .1 through .9 range test had the higher coefficient at $r_{bis.} = .90$.

Comparison of Conventional and Branching Tests. One of the branching tests was superior to any of the conventional tests for

$r_{bis.} \geq .60$ ($r_{ij} \geq .36$). At the higher biserials, .70 through .90, both of the branching tests yielded higher coefficients than did any of the conventional tests. For the lower biserials, .30 through .50, the C50 conventional tests resulted in slightly higher coefficients than did either of the branching tests.

Ten-Item Tests

Conventional Tests. The test score-ability correlation coefficients for the 10-item conventional tests are shown in the first five rows of Table 3. The data showed that the C50 test had the highest coefficient for each biserial through .60. For these same biserials, all of the .3 through .7 range tests were next highest and the .1 through .9 range tests were lowest. At $r_{bis.} = .70$, the C50 and .3 through .7 tests were about equally effective, and yielded higher coefficients than the .1 through .9 tests. At biserials of .80 and .90, the original situation was reversed and the C50 test had the lowest coefficients and the .1 through .9 tests the highest coefficients.

TABLE 3

*Test Score-Ability Correlation Coefficients for 10-Item Per Subject
Conventional and Branching Tests*

Biserial	.30	.40	.50	.60	.70	.80	.90
Test							
C (all .5)	614*	728	807	859	891	905	898
C (.3-.7, N)	608	723	802	856	890	909	910
C (.3-.7, R)	604	719	799	854	890	911	917
C (.1-.9, N)	586	702	786	844	886	913	929
C (.1-.9, R)	563	680	767	830	877	913	941
B (.3-.7, 1-PS)	612	728	808	866	904	926	931
B (.3-.7, 2-PS-E)	520	642	737	809	863	898	915
B (.3-.7, 2-PS-U)	512	633	721	799	851	885	898
B (.1-.9, 1-PS)	601	719	801	862	905	934	953
B (.1-.9, 2-PS-E)	531	655	751	825	881	921	948
B (.1-.9, 2-PS-U)	519	640	729	808	862	899	918

* Decimal points omitted.

Branching Tests. The 10-item per subject branching test data is given in the bottom six rows of Table 3. The .3 through .7 1-PS tests tended to correlate higher with the criterion than did the .1 through .9 tests through a biserial of .60. Above this level the converse held. It should be noted that for all biserials, and any given item difficulty range, the 1-PS branching test correlated higher than

any 2-PS test covering the same range. In fact, with only one exception ($r_{bis.} = .90$), both the 1-PS .1 through .9 and .3 through .7 tests yielded higher coefficients than did any of the 2-PS tests. The 2-PS-E tests correlated higher with the ability criterion than did the 2-PS-U tests.

Comparison of Conventional and Branching Tests. One of the 1-PS branching tests was superior to any of the conventional tests for biserials above .40 (the .3 through .7 1-PS was highest at $r_{bis.} = .50$ and .60; the .1 through .9 1-PS was highest at $r_{bis.} = .70$ through .90). At a biserial of .30 the C50 test coefficient was slightly higher and at $r_{bis.} = .40$ the C50 conventional and .3 through .7 1-PS branching tests had the largest coefficients. The 2-PS branching tests compared favorably with the best conventional test only at very high biserials.

Fifteen-Item Tests

Conventional. All 15-item tests were evaluated at biserials of .40, .60 and .80. The test score-ability correlation coefficients for the five conventional tests are given in the first five rows of Table 4.

TABLE 4
*Test Score-Ability Correlation Coefficients for 15-Item Per Subject
Conventional and Branching Tests*

Biserial	.40	.60	.80
Test			
C (all .50)	792 ^a	896	923
C (.3-.7, N)	787	894	928
C (.3-.7, R)	785	894	930
C (.1-.9, N)	764	884	936
C (.1-.9, R)	751	877	937
B (.3-.7, 1-PS)	793	903	943
B (.1-.9, 1-PS)	786	902	953

^a Decimal points omitted.

These data showed that the C50 test had the highest coefficient at biserials of .40 and .60. At a biserial of .80, the .1 through .9 tests (both N and R) did best. A comparison of the .3 through .7 and .1 through .9 tests across the three biserials showed that the narrower range tests received higher coefficients at the lower biserials (.40 and .60) and the wider range tests did better for the high biserial (.80). This general trend was consistent with the results obtained

for the 5- and 10-item conventional tests. For tests of a given range, those with approximately normally distributed item difficulties were superior to those with rectilinear difficulty distributions at the .40 biserial and did less well than their rectilinear counterparts at a biserial of .80. At $r_{bis.} = .60$, no difference was obtained between the .3 through .7 N and R tests, but the .1 through .9 N test was superior to the R test of the same range. This same trend was also found for the 10-item conventional tests. In general, as biserials (and thus item intercorrelations) increased, wider range tests and tests with more rectilinear item difficulty distributions did progressively better.

Branching. Data for the two 15-item branching tests are given in the last two rows of Table 4. The .3 through .7 test correlated higher with the ability criterion at the .40 biserial, while the .1 through .9 test yielded the highest coefficient at $r_{bis.} = .80$. The two branching tests were essentially equivalent at the .60 biserial.

Comparison of Conventional and Branching Tests. Both of the branching tests yielded higher coefficients than did any of the conventional tests for biserials of .60 and .80. At the .40 biserial, the C50 test was essentially equivalent to the .3 through .7 branching test.

Effects of Test Length

Table 5 gives the increments in test score-ability coefficients as the tests were increased in length from five to fifteen items. Increasing the number of items from five to ten resulted in increments in the correlations which were about twice as large as those obtained by increasing test length from 10 to 15 items. Increases in test length led to higher test score-ability coefficients for the lower biserial values. There appeared to be little difference between conventional and branching tests in terms of the effects of increasing test length.

Overview

Both conventional and branching test data showed that tests with the least spread of item difficulties yielded the highest coefficients when low to moderate biserials were assumed. For medium to high biserials, the moderate range and wide range tests tended to yield coefficients of about the same magnitude. The wide range tests gen-

TABLE 5

*Increment in Test Score-Ability Correlation Coefficients
with Increase in Test Length*

Biserial		.30	.40	.50	.60	.70	.80	.90
C (all .50)	5-10	132*	127	111	090	068	047	027
	10-15		064		037		018	
C (.3-.7, N)	5-10							
	10-15		064		038		019	
C (.1-.9, N)	5-10							
	10-15		062		040		023	
C (.3-.7, R)	5-10	131	128	113	092	071	050	030
	10-15		066		040		019	
C (.1-.9, R)	5-10	129	131	121	104	084	063	041
	10-15		071		047		024	
B (.3-.7, 1-PS)	5-10	134	129	114	092	069	046	025
	10-15		065		037		017	
B (.1-.9, 1-PS)	5-10	137	133	115	094	071	048	025
	10-15		066		039		018	

* Decimal points omitted.

erally did best when very high biserials were assumed. These data are consistent with the 9- and 18-item test data reported by Brogden (1946). The shift in the relative effectiveness of the narrower and wider range tests tended to take place earlier when test length was increased.

For the lowest biserial assumed (.30), the C50 test was the only conventional test which correlated higher with the ability criterion than did the best branching test. At biserials of .40 and .50, and 10- and 15-item branching tests covering a .3 to .7 range and the C50 test were essentially equivalent. For biserials of .60 and above, one of the branching tests always did better than any of the conventional tests.

A comparison of 1-item per stage and 2-item per stage branching tests (at the 10-item test length) indicated that the 1-item per stage tests had uniformly higher coefficients than did the 2-item per stage tests of the same range. In view of these results, it would not seem profitable to use the more complex 2-item per stage structure in the development of branching tests for the purpose of maximizing overall correlation.

REFERENCES

- Bayroff, A. G., Thomas, J. A., and Anderson, A. A. Construction of an experimental sequential item test. U. S. Army Personnel Research Office. Research Memorandum 60-1. January 1960.

- Brogden, H. E. Variation in test validity with variation in the distribution of item difficulties, number of items, and degree of their intercorrelation. *Psychometrika*, 1946, 11, 197-213.
- Krathwohl, D. R. and Huyser, R. J. The sequential item test (SIT). *American Psychologist*, 11, 1956, 419.
- Lord, F. M. A theory of test scores. *Psychometric Monograph*, No. 7, 1952.
- Seeley, L. C., Morton, Mary A., and Anderson, A. A. Exploratory study of a sequential item test. U. S. Army Personnel Research Office. Technical Research Note 129. December 1962.
- Tucker, L. R. Maximum validity of a test with equivalent items. *Psychometrika*, 1946, 11, 1-3.
- Waters, Carrie J. Preliminary evaluation of simulated branching tests. U. S. Army Personnel Research Office. Technical Research Note 140. June 1964.

MINIMIZING ORDER EFFECTS IN THE SEMANTIC DIFFERENTIAL¹

ROBERT B. KANE
Purdue University

HOUSTON (1967) has devised a theoretical solution to the problem of order effects on responses in tests or questionnaires of k items, for $k \leq 22$, but there is no reasonable way to utilize these results if ordinary duplicating equipment is used to prepare the materials. Kane (1969) provided a computer program that generates semantic differential (SD) questionnaires which takes account of all three sources of presentation order effects in a SD: (1) concept presentation order; (2) adjective scale order; and (3) the polarity of each scale (which end is positive). Bias due to concept order and scale order are minimized by using the particular permutation of k items found by Houston to yield the minimum order effect. Scale polarity is determined scale-by-scale by reference to a pseudo-random digit generating function. The program was written so that E may invoke or ignore subroutines designed to minimize each type of order effect. For example, E may minimize bias due to scale ordering while holding concept order and scale polarity invariant.

Experimental Design

Nine SD questionnaire generating strategies, eight combinations producible from the computer program plus the standard noncomputer based format, were considered.

Of the 36 pairs of strategies derivable from these nine, three were compared to determine the utility of reducing order effects when

¹The work reported herein was performed pursuant to a grant from the Office of Education, U. S. Department of Health, Education, and Welfare.

employing the *SD*. The study of each pair is designated as an experiment.

Experiment I: Concept order, scale order, and scale polarity fixed vs. all three varied.

These strategies should produce maximum differences with respect to order effects.

Experiment II: A few concept orders, scale order and polarity fixed vs. all three varied.

These strategies should produce differences comparable to those between *SD* questionnaires produced in the standard (non-computer-based) way and those produced by employing all the format variability available from the computer.

Experiment III: A few concept orders, scale order and polarity fixed vs. concept and scale order fixed, polarity varied.

This provides a comparison between the noncomputerized questionnaire and one in which only scale polarity is varied. If significant differences in response patterns are found within Experiments II and III, and if the differences in Experiment II and III are comparable then it would be economically sound to generate *SD* questionnaires varying only scale polarity since it is simpler (thus less costly) than varying all three orders.

One hundred fifty undergraduate elementary education majors were selected randomly as *Ss*. Fifty *Ss* were assigned randomly to each experiment. Within each experiment 25 *Ss* were assigned randomly to each treatment. Ten days after the first data collection each *S* completed another *SD* composed of the same concepts and adjective scales, but generated by the opposing strategy.

Each *SD* questionnaire was composed of nine concepts related to major curricular areas in the elementary schools each to be rated on 14 adjectival scales.

Findings and Analysis

Fifty-four (two treatments \times 3 experiments \times 9 concepts) 14×14 matrices of product-moment correlations were computed. Each was factored and rotated to the Varimax criterion (Kaiser, 1958, 1960). Since the proportion of total variance accounted for by the first two factors ranged from 0.45 to 0.82, and since factor III contributed more than 10 per cent of the total variance in only

one case, only factors I and II were used as data sources for this study.

Three comparisons between responses to the two types of *SD* questionnaires were analyzed in each experiment:

1. Differences in rotated factor structure.
2. Differences in factor scores concept-by-concept.
3. Differences in response consistency.

Factor Structure

Scales with factor loadings ≥ 0.30 were listed for factors I and II for each of the 54 rotated factor matrices. In the case of factor I these data then were compressed by recording only those scales with loadings ≥ 0.30 for eight concepts out of nine. In the case of factor II the criterion for final recording of a scale was set at loadings ≥ 0.30 for seven concepts out of nine. Tables 1 and 2 list the scales which survived these screenings.

In Table 1 there are 56 (i.e., 4 strategies \times 14 scales) cells in which a tally mark can appear. By changing the entry in just six of these cells identical matchings could be created in all four

TABLE 1

Scales With Factor I Loadings ≥ 0.30 In At Least 45 Out of 54 Cases

	Questionnaire Generating Strategy			
	1 All Orderings Fixed	2 All Orderings Varied	3 Only Concept Order Varied	4 Only Scale Polarity Varied
heavy-light			X	X
active-passive	X		X	X
happy-sad ^a	X	X		X
heavenly-hellish	X	X		
fast-slow	X		X	X
positive-negative ^a	X	X		
difficult-easy			X	X
optimistic-pessimistic	X		X	X
strong-weak ^a	X	X		
hard-soft			X	X
nice-awful ^a	X	X	X	
hot-cold	X		X	
good-bad ^a	X	X	X	X
masculine-feminine				

^a Denotes scales chosen to represent factor I in factor score study.

TABLE 2

Scales With Factor II Loadings ≥ 0.30 In At Least 36 Out of 54 Cases

	Questionnaire Generating Strategy			
	1	2	3	4
	All Orderings Fixed	All Orderings Varied	Only Concept Order Varied	Only Scale Polarity Varied
heavy-light*	X	X	X	X
active-passive				
happy-sad				
heavenly-hellish	X			
fast-slow				
positive-negative				
difficult-easy*	X	X	X	X
optimistic-pessimistic				
strong-weak				
hard-soft*	X	X	X	X
nice-awful	X			
hot-cold				
good-bad				
masculine-feminine				

* Denotes scales chosen to represent factor II in factor score study.

strategy columns. Indeed, identical markings already exist for nine of the 14 scales. Although the strategy two column exhibits the greatest deviation from the other columns, these data indicate marked similarities among the columns.

By changing only two entries out of 56 in Table 2 matchings in all four strategy columns could be created. With the possible exception of factor I, strategy 2, there seem to be no appreciable differences in factor structure among the four questionnaire generating strategies for either factor I or factor II.

Factor Scores

Five scales were chosen to represent factor I and three scales were chosen to represent factor II.

A score from 0 to 6 was recorded for each *S* on each scale and mean scores of factor I scales as well as factor II scales were computed concept-by-concept within each experimental treatment. Thus within each experiment there were nine pairs of mean scores for factor II. The difference between mean scores within each pair was analyzed by an analysis of variance model.

Table 3 lists the F ratios emanating from experiments I, II, and III respectively.

TABLE 3
F Ratios for Factor Score Study

Concept	Experiment I		Experiment II		Experiment III	
	Factor I	Factor II	Factor I	Factor II	Factor I	Factor II
Language Arts	1.810	0.844	0.258	0.067	0.064	0.418
Mathematics	1.881	1.715	0.508	0.197	0.039	0.146
Science	0.051	0.717	0.646	1.449	0.602	0.054
Social Studies	2.843 ^a	0.524	1.114	0.829	0.050	0.336
Teaching Children	0.746	1.246	2.173	2.767 ^a	0.008	0.803
Teaching Children Language Arts	0.346	0.280	0.580	1.294	0.330	1.200
Teaching Children Mathematics	0.945	0.000	0.256	0.375	0.788	2.863 ^a
Teaching Children Science	0.026	0.006	0.007	0.401	0.002	1.126
Teaching Children Social Studies	0.377	0.006	0.098	0.181	0.792	1.774

^a Significant at $\alpha = 0.10$. None of these F ratios is significant at $\alpha = 0.05$.

Of the 54 F ratios displayed in Table 3, none is significant at the $\alpha = 0.05$ level; only three are significant at the $\alpha = 0.10$ level. In fact only six more are significant when the α level is advanced to 0.25. Forty of the 54 F ratios are less than 1.000. These data indicate that no systematic differences in factor scores occur in any of the experiments.

Response Consistency

As a final reading of the differences between strategies a measure designated "response consistency" was devised. This measure seeks to answer the following question: How closely does S 's response on the $(n + k)$ th adjective scale conform to his response on the n th adjective scale? To answer this question the absolute value of the difference between the score on scale n and scale $n + k$ was selected as the measure. Thus

$$C = \frac{\sum |s_n - s_{n+k}|}{14 - k},$$

where C denotes a response consistency index, s_n denotes the score on scale n , s_{n+k} denotes the score on scale $(n + k)$, and $|s_n -$

s_{n+k} is summed over all such differences within a given concept. Clearly, the summation could be made of all such differences produced by a given S across concepts if one wished to do so. Summing within concepts and across S s was done to conform with the other analyses made in this study. It was decided to let $k = 1, 2, 3$, or 4. By using $14-k$ in the denominator response consistency measures for $k = 1, 2, 3$, and 4 were transformed into comparable indices.

If order effects are salient then differences when $k = 1$ should be less than differences when $k = 2$ and, in general $C_1 < C_2 < C_3 < C_4$. Table 4 displays response consistency indices for $k = 1, 2, 3$, and 4 for each concept from 200 of the *SD* questionnaires completed by S s in this study.

An inspection of the nine columns of Table 4 does not support the existence of the order relation.

$$C_1 < C_2 < C_3 < C_4.$$

In order to determine whether row or columnar differences are significant a two way analysis of variance was performed. Table 5 includes the relevant data.

The inequality of the indices across concepts suggests that the magnitude of the response consistency indices is related to the concept being rated. While the F ratio associated with within-column differences suggests that $C_1 \neq C_2 \neq C_3 \neq C_4$, no systematic order relation among C_1, C_2, C_3 , and C_4 , was observed. There is no evidence of concept-index interaction. Thus while there are differences in response consistency these differences do not appear to be interpretable as indicators of order effects.

This research supplied no evidence that users of the *SD* need to be concerned about item order effects as a significant source of error variance. In Experiment I, where one treatment invited maximum order effects and the other treatment minimized the sources of these effects no significant response differences were observed. In Experiments II and III, where the opposing *SD* formats were less profoundly different, the same result obtained.

Subject to the usual constraints on the generalizableness of findings it appears the E s may cease worrying about the effects of a constant item presentation order when administering the *SD*.

TABLE 4
Response Consistency Indices Based on 200 SD Questionnaires

<i>k</i>	Language Arts	Mathematics	Science	Social Studies	Teaching Children	Teaching Children Language Arts	Teaching Children Mathematics	Teaching Children Science	Teaching Children Social Studies
1	356.2	356.8	266.2	332.7	399.2	355.1	312.2	313.1	328.7
2	331.6	364.3	274.5	317.5	339.1	317.3	311.8	304.7	294.0
3	328.2	342.0	266.4	299.4	348.6	318.0	312.1	297.8	302.8
4	358.3	357.1	272.3	331.6	351.3	348.0	316.6	315.7	328.5

TABLE 5

*A Comparison of Four Response Consistency Indices Across Nine
SD Concepts for 200 SD Questionnaires*

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Response Consistency Indices for $k = 1, 2, 3$, and 4	14.12	3	4.71	10.5*
Concepts	116.58	8	14.57	32.4*
Interaction	11.01	24	0.46	1.0
Within Cells	3251.88	7164	0.45	

* Significant at $\alpha = 0.01$.

REFERENCES

- Houston, T. R. A source of artifact in multiple response designs. Paper presented at the annual convention of the American Educational Research Association, New York, February, 1967.
- Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 1958, 23, 187-200.
- Kaiser, H. F. The application of electronic computers to factor analysis. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1960, 20, 141-151.
- Kane, R. B. Computer generation of semantic differential questionnaires. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 191-192.

BEHAVIORAL COGNITION AS RELATED TO INTERPERSONAL PERCEPTION AND SOME PERSONALITY TRAITS OF COLLEGE STUDENTS¹

C. M. N. MEHROTRA

The Ohio State University²

THE Structure of Intellect(SI) Model proposed by Guilford (1967) classifies the intellect into operations which it can perform, different contents of these operations, and different products. The model hypothesizes four kinds of content—semantic, symbolic, figural, and behavioral. The first three content areas are not unfamiliar to the psychologists, as they were included previously in other models, for example, Eysenck (1953), and are seen in most of the existing test batteries. The fourth content area, the behavioral one, has been added to the model "to take care of operations pertaining to the behavior of other people" (Guilford, 1967). This area includes feelings, motives, thoughts, intentions or other psychological dispositions which might affect an individual's social behavior. O'Sullivan, Guilford, and de Mille (1965) have developed measures for six different factors of behavioral cognition hypothesized in the SI model and have shown that these abilities are factorially distinct from previously isolated *intellectual* abilities. But there is no research evidence to show whether the behavioral cognition factors are different from the related factors in the nonintellectual domain. The present study was conducted to achieve this goal. To be more specific, the aim was to determine the nature of the relationship

¹ This paper is based on a dissertation submitted to The Ohio State University in partial fulfillment of the requirements for the Ph.D. degree. The author wishes to thank John E. Horrocks, Chairman of his Committee, George G. Thompson, Robert J. Wherry, Phillip M. Clark, and Edwin G. Novak for their advice and criticism.

² Now at the Educational Testing Service.

of behavioral cognition factors with interests (social service, persuasive, artistic, and literary), values (social, political, and aesthetic) and personality variables (inclusion, control, affection, extraversion-introversion, sensing-intuition, judgment-perception, and thinking-feeling). It was hypothesized that if the behavioral cognition measures developed by O'Sullivan et al., (1965) have discriminant validity they will have low relationship with these variables.

Method

Subjects

The subjects were 100 male and 100 female college undergraduates enrolled in an introductory course in educational psychology at The Ohio State University.

Psychological Measures

The tests used in the present study were of behavioral content involving the operation of cognition and the products of classes, systems, transformations, and implications. Thus the following tests were selected: Expression Grouping for cognition of behavioral classes (CBC), Missing Cartoons for cognition of behavioral systems (CBS), Social Translations for cognition of behavioral transformation (CBT), and Cartoon Predictions for cognition of behavioral implications (CBI). A description of the four factors and their tests can be found in Hoepfner and O'Sullivan (1968). For assessment of interpersonal perception six filmed interviews (Cline and Richards, 1960) were used. Three of these interviewees were male and three were female. After viewing each film the Ss were required to fill out the three judging instruments: CPI Opinion Prediction Test (OPT), Adjective Checklist (ACL) and Behavior Postdiction Test (BPT). The variables in the nonintellectual domain were assessed by self-descriptive, objectively scored inventories. As far as possible two instruments were employed for measuring the variables in one area. Personality variables were assessed by using the FIRO-B (Schutz, 1967) and Myer-Briggs Type Indicator (Myers, 1962), interests by the Kuder Preference Record (Kuder, 1942) and values by the Study of Values (Allport, Vernon and Lindzey, 1960). Rationale for choosing each of these instruments is given in Mehrotra (1968).

Correlations were computed to separately examine the relationship between each of the behavioral cognition scores and the interpersonal perception, personality, interests, and values scores. Multiple regression analysis was used to see how far it is possible to predict the performance on behavioral cognition tests by using the weighted sum of scores on measures of interpersonal perception, personality, interests and values.

Results

Table 1 contains the intercorrelations among the five scores on measures of behavioral cognition for the total sample.³ All of these 10 coefficients are significant at .01 level. This indicates that the four behavioral cognition factors are not statistically independent of one another. The Social Translations Test (CBT), which is the only test using verbal material, has the lowest correlation with other measures of behavioral cognition. These high correlations can be interpreted as indicative of mutual involvement of abilities in solving the problems with behavioral content. In terms of the criteria specified by Campbell and Fiske (1959), one might say that as these measures are conceptually independent (Guilford, 1964), the high correlations are indicative of inadequate discriminant validity of these measures. However, in view of the fact that their construct validity has already been established by O'Sullivan et al., (1965), these correlations may be interpreted as reflecting the at-

TABLE 1
*Intercorrelations among the Five Behavioral Cognition Scores
for the Total Sample*

Tests	2	3	4	22	M	SD
1. Missing Cartoons (CBS)	457*	341*	550*	770*	19.405	4.736
2. Expression Grouping (CBC)		403*	553*	747*	19.075	3.846
3. Social Translations (CBT)			523*	696*	17.100	4.188
4. Cartoon Prediction (CBI)				838*	22.075	4.073
22. Composite					77.710	13.155

Note.—* Significant at $p = .01$ level, nondirectional test. Decimal points are omitted from correlations only.

³ Intercorrelations were also computed for the male and the female samples separately. Though there was a sex difference in the magnitude of r 's, there was no evidence for a sex difference in the patterning of the coefficients. Correlating the matched r 's across sex had a Spearman rho of .70 which is statistically significant ($p < .05$).

tributes of the testees and not as indicators of only the formal properties of the tests, as implied by the concept of discriminant validity (Kroger, 1968).

Table 2 shows the relationship of measures of behavioral cognition with those of interpersonal perception. The Cartoon Predictions Test does not have a significant correlation with Opinion Prediction and Behavior Postdiction measures. However, its correlation with Adjective Checklist F (ACLF) is statistically significant, but that is also true for the correlations of other behavioral cognition measures with ACLF. When the correlations of behavioral cognition scores and ACLF were obtained separately for the male and female sample, they were significant only in the male sample.

It may be recalled that the SI product category of implications is concerned with extrapolations from given information to either its antecedents or its consequents. Cartoon Predictions, the measure of CBI, is based mainly on the consequent part of the definition. The measures of interpersonal perception are also concerned with the extrapolations from given information. The main difference between

TABLE 2

Correlations of Behavioral Cognition Scores and Interpersonal Perception Scores for the Total Sample (N = 200)

Interpersonal Perception	Behavioral Cognition				22. Composite
	1. Missing Cartoons (CBS)	2. Expression Grouping (CBC)	3. Social Translation (CBT)	4. Cartoon Predictions (CBI)	
26. Opinion Prediction M	-.104	-.073	-.058	-.065	-.082
30. Adjective Checklist M	.136	.109	.117	.111	.134
34. Behavior Postdiction M	-.140 ^a	-.079	-.043	-.121	-.127
35. Composite M	-.022	.002	.030	-.015	-.008
39. Opinion Prediction F	-.020	.014	.027	.009	.005
43. Adjective Checklist F	.174 ^a	.205 ^b	.240 ^b	.170 ^a	.242 ^b
47. Behavior Postdiction F	-.166 ^a	.000	.030	-.106	-.090
48. Composite F	.010	.114	.165 ^a	.046	.102

^a Significant at $p = .05$ level, nondirectional test.

^b Significant at $p = .01$ level, nondirectional test.

Note.—Decimal points have been omitted.

these two sets of instruments is that of method: filmed interviews versus immobile photographs and cartoons. Since convergent validity is represented in the agreement between two attempts to measure the same trait through maximally different methods, it may be said that the data in the present study failed to provide enough evidence for convergent validity of Cartoon Predictions Test. It is possible that when immobile photographs, cartoons and drawings of faces are used, one may be measuring a variable which is different from what is being measured by using a filmed interview which provides a number of other cues e.g., tone, interaction with other persons, behavior in different situations. Other possible reasons for these low relationships include (a) low reliability of OPT, ACL and BPT which were used to assess accuracy of interpersonal prediction, (b) low reliability of Cartoon Predictions Tests which was used as a measure of CBI, and (c) the restricted range of scores on each of the measures employed in the present study.

Table 3 shows the relationship of behavioral cognition scores with personality traits. Missing Cartoons, the CBS measure used in this study, has a statistically significant correlation with Sensing-Intuition and Judging-Perceiving Scales of the MBTI. This indicates that those who obtain high scores on CBS also tend to do well on (a) sensing (being aware of things directly through one or another of five senses) as opposed to intuition and (b) judging (coming to a conclusion by shutting off perception for the time being) as opposed to perception. The Expression Grouping (CBC) has a statistically significant correlation only with the Extraversion-Introversion Scale of MBTI; this shows that extraverted Ss tend to do well on CBC. Social Translations (CBT) is not correlated with any of the personality variables, and Cartoon Predictions (CBI) is correlated only with Sensing-Intuition.

Table 4 contains the correlations of behavioral cognition scores with interests and values. Both Missing Cartoons (CBS) and Cartoon Predictions (CBI), the measures using cartoons, correlate significantly to artistic interests. This indicates that Ss who enjoy artistic activities tend to obtain high scores on measures of CBS and CBI, which in the present study use cartoons. Further studies are needed to see if the artistic interests continue to be correlated with CBS and CBI even when the measures do not use cartoons

TABLE 3

Correlations between Behavioral Cognition Scores and Scores on Personality Variables for the Total Sample (N = 200)

Personality Variable	Behavioral Cognition				5. Composite
	1. Missing Cartoons (CBS)	2. Expression Grouping (CBC)	3. Social Translation (CBT)	4. Cartoon Predictions (CBI)	
18. Extraversion-Introversion	107	155 ^a	096	040	113
19. Sensing-Intuition	255 ^b	041	124	168 ^a	198 ^b
20. Thinking-Feeling	027	-031	081	098	080
21. Judging-Perceiving	189 ^b	033	107	054	101
50. Expressed Inclusion	-015	-020	057	110	033
51. Wanted Inclusion	-116	011	047	013	-023
52. Expressed Affection	-070	018	111	127	050
53. Wanted Affection	-020	020	086	077	046
54. Expressed Control	001	-042	-115	013	-046
55. Wanted Control	-101	-010	-058	008	-045

^a Significant at $p = .05$ level, nondirectional test.

^b Significant at $p = .01$ level, nondirectional test.

Note.—Decimal points have been omitted.

and drawings. Negative correlations of CBT and CBI with persuasive interests show that those who enjoy persuasive activities may not do well on tasks involving behavioral transformation or extrapolation. When the data was analyzed separately for the male

TABLE 4

Correlations of Behavioral Cognition Scores with Scores on Measures of Interests and Values for the Total Sample (N = 200)

Interests and Values	Behavioral Cognition				22. Composite
	1. Missing Cartoons (CBS)	2. Expression Grouping (CBC)	3. Social Translation (CBT)	4. Cartoon Predictions (CBI)	
8. Persuasive Interests	-095	-111	-202 ^b	-163 ^a	-168 ^a
9. Artistic Interests	318 ^b	084	090	193 ^b	214 ^b
10. Literary Interests	077	045	034	-011	057
12. Social Service Interests	-115	042	121	-033	013
14. Economic Values	-052	-061	-151	-100	-114
15. Aesthetic Values	125	-021	054	024	062
16. Social Values	-157 ^a	-065	059	-077	-067
17. Political Values	-010	041	-095	020	-022

Note.—Decimal points have been omitted.

^a Significant at $p = .05$ level, nondirectional test.

^b Significant at $p = .01$ level, nondirectional test.

TABLE 5
Variables with Significant Partial F Values in the Multiple Regression Analyses for the Four Behavioral Cognition Factors

Dependent Variable	Girls (N = 100)			Boys (N = 100)			Total Sample (N = 200)		
	Independent Variables	R	R ²	Independent Variables	R	R ²	Independent Variables	R	R ²
Missing Cartoon (CBS)	Judgment-Perception Artistic Interests Behavior Postdiction F Adjective Checklist F Extraversion-Introversion	.4725*	.2233	Artistic Interests Behavior Postdiction M Adjective Checklist M Extraversion-Introversion Sensing-Intuition Social Values	.6027*	.3632	Artistic Interests Sex of the Student Sensing-Intuition Adjective Checklist M Extraversion-Introversion Behavior Postdiction M Adjective Checklist F Expressed Inclusion	.5156*	.2659
Expressions Grouping (CBC)	Extraversion-Introversion Adjective Checklist F Opinion Prediction F	.3681*	.1355	Adjective Checklist F Adjective Checklist M Wanted Control	.3872*	.1499	Extraversion-Introversion Behavior Postdiction M Sensing-Intuition Social Service Interests	.3381*	.1143
Social Translations (CBT)	Extraversion-Introversion Sensing-Intuition Political Values	.3447*	.1188	Adjective Checklist F Persuasive Interests Social Service Interests	.4176*	.1744	Persuasive Interests Social Service Interests	.2354*	.0554
Cartoon Prediction (CBI)	Social Service Interests Opinion Prediction M Thinking-Feeling Adjective Checklist M Behavior Postdiction F	.5068*	.2568	Persuasive Interests Expressed Inclusion Wanted Control	.3871*	.1499	Adjective Checklist F Persuasive Interests Behavior Postdiction F	.3576*	.1279

* These values of multiple correlation coefficient (R) were obtained at the end of the last step where the additional predictors entering into the equations were significant. In each case the Multiple R was significant at .01 level.

and female sample, some sex differences were found in the pattern of relationships.

Multiple regression analysis was performed for each dependent variable by using the Stepwise procedure. Table 5 shows the independent variables with significant partial F values. Although the multiple correlations for the four dependent variables were statistically significant at the .01 level, different variables accounted for the prediction in every regression equation. Except in the case of Missing Cartoons (CBS), a different set of predictors was utilized in the prediction of the dependent variable in the male and female samples. In the female sample Extraversion-Introversion had a significant partial F value in predicting the performance on CBS, CBC, and CBT measures, while in the male sample it was significant only in the prediction of CBS scores.

Summary

This study was designed to determine the relationship of behavioral cognition with interpersonal perception and personality traits. Though cognition of behavioral implications is by definition very similar to interpersonal prediction, the present study did not find significant correlations between them. As two different methods were used to measure these variables, this was considered an indication of inadequate convergent validity. The pattern of relationship of behavioral cognition factors with the personality variables was different in the male and female samples. In the total sample CBS had statistically significant correlation with Sensing-Intuition and Judging-Perceiving scales of MBTI, CBC with Extraversion-Introversion and CBI with Sensing-Intuition. Both Missing Cartoons (CBS) and Cartoon Predictions (CBI), the measures using cartoons, correlated significantly with artistic interests. Significant multiple correlations were obtained for each of the dependent variables.

REFERENCES

- Allport, G. W., Vernon, P. E., and Gardner, L. *Study of values: A scale for measuring the dominant interests in personality*, third edition. Boston: Houghton Mifflin Co., 1960.
- Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.

- Cline, V. S. and Richards, J. M., Jr. Accuracy of interpersonal perception—A general trait? *Journal of Abnormal and Social Psychology*, 1960, 60, 1-7.
- Eysenck, H. J. *Uses and abuses of psychology*. London: Pelican, 1953.
- Guilford, J. P. Zero intercorrelations among tests of intellectual abilities. *Psychological Bulletin*, 1964, 61, 401-404.
- Guilford, J. P. *The nature of human intelligence*. New York: McGraw-Hill, 1967.
- Hoepfner, R. and O'Sullivan, M. Social intelligence and IQ. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 339-344.
- Kroger, R. O. Conceptual and empirical independence in test validation: A note on Campbell and Fiske's discriminant validity. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 383-387.
- Kuder, G. F. *Kuder Preference Record (Vocational) Form BM*. Chicago: Science Research Associates, 1942.
- Mehrotra, C. M. N. Behavioral cognition as related to interpersonal perception and some personality traits of college students. (Unpublished Ph.D. dissertation, The Ohio State University), 1968.
- Myers, I. B. *The Myers-Briggs type indicator*. Princeton, New Jersey: Educational Testing Service, 1962.
- O'Sullivan, M., Guilford, J. P. and deMille, R. *The measurement of social intelligence*. Reports from the Psychological Laboratory, No. 34, Los Angeles: The University of Southern California, 1965.
- Schutz, W. C. *FIRO-B*. Palo Alto: Consulting Psychologists Press, 1967.

VOCATIONAL INTERESTS AND INTELLIGENCE IN GIFTED ADOLESCENTS

GEORGE S. WELSH

The University of North Carolina at Chapel Hill

In a previous study of gifted adolescents it was shown that verbal interests and differential performance on a verbal and a nonverbal intelligence test were significantly related (Welsh, 1967). Verbal interest was inferred from scores on three scales of the Strong Vocational Interest Blank (SVIB) (Strong, 1959); Advertising Man, Lawyer, and Author-Journalist. Verbal intelligence was measured by Terman's Concept Mastery Test (CMT) (Terman, 1956) and nonverbal intelligence by the D-48 (Black, 1963). Subjects with greater verbal interest scored relatively higher in verbal intelligence and, conversely, subjects higher in verbal intelligence showed higher scores on verbal interest scales.

The present report is a further study of the same subjects using a correlational analysis of SVIB scales with the CMT and the D-48. In addition to CMT total-score, separate part-scores for the two sections, Vocabulary (Synonyms-Antonyms) and Analogies, were also utilized since it has been found that for these subjects Analogies is more highly correlated with the D-48 than is Vocabulary (Welsh, 1969).

From the SVIB 55 regular vocational and nonvocational scales were examined plus four newly developed scales. These special scales resulted from an item analysis of subgroups of the gifted adolescents scoring relatively high or low on a figure preference test art scale, often used as an index of creative potential (Welsh, 1959), conjointly with high or low scores on the CMT.

These scales have been conceptualized along two independent dimensions. The first dimension, now called "origence," contrasts

those at the low end who prefer an organized, well-structured, obvious, and explicit situation versus the high origent person who is more at home in an open, diffuse, subtle, and implicit task. The second dimension, "intellectence," contrasts those at the low end who seem to favor concrete and literal experience versus the abstract-conceptual approach of the high intellectent person. These dimensions have been developed in a two-factor model of creativity (Welsh, in press).

The following nomenclature has been adopted for the four special SVIB scales:¹

S-1	S-2
High origence	High origence
Low intellectence	High intellectence
S-3	S-4
Low origence	Low origence
Low intellectence	High intellectence

Table 1 gives the correlations of all 59 SVIB scales with intelligence test scores for the sexes separately.

TABLE 1
Correlations of Strong Scales with Intelligence Test Scores

	Male				Female			
	Terman		CMT		Terman		CMT	
	Total	Vocab	Anal	D-48	Total	Vocab	Anal	D-48
I								
Artist	17	16	16	03	17	20	11	-04
Psychologist	47	43	46	24	41	38	38	00
Architect	19	16	19	11	25	25	22	00
Physician	29	23	34	20	31	29	29	22
Psychiatrist	35	32	33	14	32	30	30	22
Osteopath	-02	-04	02	-01	06	06	05	10
Dentist	02	-03	09	02	06	06	04	00
Veterinarian	-25	-27	-17	-06	-16	-16	-13	-00
II								
Mathematician	39	31	44	32	30	26	30	22
Physicist	29	21	35	27	25	22	26	22
Chemist	31	23	38	27	31	27	31	22
Engineer	16	09	24	21	21	17	23	22

¹ These special scales may be scored on the current form of the SVIB, 399 T, by arrangement with Prediction, Inc., Box 298, Greensboro, N. C. 27402; or write directly to National Computer Systems, 1015 South Sixth St., Minneapolis, Minn. 55415.

TABLE 1—Continued

	Male				Female			
	Terman		CMT		Terman		CMT	
	Total	Vocab	Anal	D-48	Total	Vocab	Anal	D-48
III								
Production Manager	-17	-18	-14	03	-16	-17	-11	06
IV								
Farmer	-09	-13	-02	06	04	02	06	07
Carpenter	-24	-25	-18	-02	-15	-16	-11	01
Forest Service Man	-08	-11	-02	06	08	06	08	08
Aviator	02	-02	07	14	19	18	18	18
Printer	02	00	04	07	15	14	14	15
Math-Science Teacher	06	01	13	20	14	09	20	26
Industrial Arts Teacher	-20	-22	-12	02	-03	-07	02	10
Voc-Agriculture Teacher	-18	-21	-10	04	-08	-11	-02	02
Policeman	-28	-27	-25	-14	-21	-20	-19	-03
Army Officer	08	06	10	11	22	18	23	23
V								
TMCA Physical Director	-18	-16	-18	-09	-09	-09	-08	03
Personnel Manager	03	05	-01	-03	05	05	05	05
Public Administrator	21	19	20	08	19	16	20	14
Vocational Counselor	01	02	-02	-01	01	-01	03	07
Physical Therapist	-03	-04	00	05	07	06	08	17
Social Worker	12	15	06	-03	12	12	10	07
Social Science Teacher	-04	-01	-08	-09	-01	-02	-01	-01
Business-Educ. Teacher	-08	-06	-09	-03	-03	-06	01	10
School Supt.	14	15	11	06	09	07	10	03
Minister	13	14	09	00	12	12	10	06
VI								
Musician	11	12	08	06	17	18	12	07
Music Teacher	02	05	-03	-02	05	06	03	02
VII								
CPA Owner	32	31	28	13	09	07	10	12
VIII								
Senior CPA	09	05	13	19	11	05	17	28
Accountant	-08	-08	-07	02	-14	-17	-07	11
Office Worker	-28	-25	-28	-12	-21	-23	-15	03
Credit Manager	-09	-07	-11	-02	-03	-06	01	13

TABLE 1—Continued

	Male				Female			
	Terman		CMT		Terman		CMT	
	Total	Vocab	Anal	D-48	Total	Vocab	Anal	D-48
Purchasing Agent	-.32	-.29	-.30	-.09	-.28	-.29	-.22	-.18
Banker	-.30	-.25	-.31	-.14	-.37	-.37	-.30	-.18
Pharmacist	-.34	-.32	-.31	-.13	-.25	-.26	-.19	-.08
Mortician	-.51	-.43	-.52	-.22	-.42	-.39	-.38	-.18
IX								
Sales Manager	-.25	-.17	-.31	-.15	-.26	-.24	-.26	-.18
Real Estate								
Salesman	-.29	-.21	-.36	-.23	-.25	-.21	-.27	-.22
Life Insurance								
Salesman	-.27	-.19	-.33	-.24	-.30	-.25	-.32	-.22
X								
Advertising Man	.04	.10	-.04	-.12	.05	.09	-.02	-.18
Lawyer	.27	.28	.21	.00	.13	.16	.07	-.08
Author-Journalist	.22	.22	.17	-.02	.15	.19	.08	-.18
XI								
President, Mfg.								
Concern	-.10	-.08	-.11	-.09	-.23	-.22	-.20	-.18
Non-Occup.								
Specialization								
Level	.43	.41	.39	.18	.30	.28	.29	.18
Interest Maturity	.07	.08	.05	-.01	.12	.09	.13	.08
Occupational Level	.24	.22	.22	.09	.12	.12	.12	.08
Masculinity	.02	-.03	.08	.13	.12	.09	.14	.18
New								
S-1	-.44	-.35	-.49	-.29	-.41	-.34	-.44	-.38
S-2	.42	.42	.36	.09	.44	.44	.37	.18
S-3	-.37	-.35	-.33	-.15	-.39	-.38	-.34	-.08
S-4	.43	.34	.48	.41	.31	.24	.36	.38

Note.—Decimal points preceding entries have been omitted. Size of groups and significance levels for r as follows:

	Male				Female	
	N	Male		N	Female	
		CMT	D-48		CMT	D-48
		529	350		632	419
Significance	.05	.085	.105		.078	.096
Level	.01	.112	.138		.103	.126

It was anticipated from the previous study that interests would be more highly correlated with the CMT than with the D-48. This is well confirmed by the number of coefficients of correlation significant beyond the .05 level. This is true for both males and females. For the sexes combined, 89 or 76 per cent of the total-score CMT correlations are significant, while only 52 per cent of the

D-48 r 's reach that level. The frequencies of significant and insignificant correlations yield a chi-square of 30.25 with $p < .0005$. The total numbers of significant and insignificant coefficients are summarized in Table 2.

It may be seen that females show relatively more significant positive correlations for both tests and particularly for the D-48. Chi-square analysis, however, indicates that this trend is insignificant.

Examination of the verbal-linguistic scales of Group X employed in the previous study shows that although Lawyer and Author-Journalist are significantly correlated with total CMT scores, Ad-

TABLE 2
Summary of SVIB Scale Correlations with Total CMT and D-48 Scores

Correlation	CMT			D-48		
	Male	Female	Total	Male	Female	Total
positive	22	29	51	16	25	41
negative	20	18	38	11	9	20
Significant*	42	47	89	27	34	61
Insignificant	17	12	29	33	25	57

* At or beyond .05 level.

vertising Man is not. This latter scale is, however, significantly negatively correlated with the D-48. When difference in magnitude and in direction of the CMT and the D-48 correlations are considered together, the pattern of r s for all three scales is consistent in showing an association of verbal interests and the CMT. Using a different method of analysis, then, some confirmation was obtained for the previous report of systematic relationship between verbal interests and verbal intelligence scores.

As mentioned above, the correlation of the D-48 and CMT part-scores was higher for Analogies than for Vocabulary; the actual correlations are .49 and .33 respectively. It would follow that SVIB scales more highly correlated with the D-48 than the CMT should have a pattern of correlation with CMT part-scores in which the coefficient for Analogies is higher than that for Vocabulary. Interest scales more highly correlated with the CMT should show a pattern of Vocabulary greater than Analogies. An analysis of these correlational patterns is given in Table 3.

TABLE 3

Summary of CMT Vocabulary-Analogies Correlational Pattern for Intelligence Tests and Interest Scales

	Vocab \geq Anal			Anal $>$ Vocab			
Correlation	Male	Female		Male	Female		Totals
Significant on CMT only							
Positive*	9	7	16	1	2	3	
Negative	6	10	16	4	0	4	
			<hr/>			<hr/>	
			32			7	39
Significant on D-48 only							
Positive	0	1	1	4	4	8	
Negative	0	0	0	1	1	2	
			<hr/>			<hr/>	
			1			10	11
Significant on CMT and D-48							
Positive	2	7	9	10	13	23	
Negative	3	4	7	7	4	11	
			<hr/>			<hr/>	
			16			34	50
Insignificant							
	8	4	12	4	2	6	18
			<hr/>			<hr/>	
Totals			61			57	118

* Includes one case each with r positive on one test, negative on other.

There were 39 interest scales significantly correlated with the CMT but not with the D-48; of these, 32 showed a pattern of Vocabulary equal or greater than Analogies while only seven showed the reverse relationship. Chi-square analysis gives a value of 16.00 with $p < .0005$.

Of the 11 interest scales correlated with the D-48 but not the CMT, 10 showed a pattern of Analogies greater than Vocabulary. The chi-square of 9.32 has $p < .005$.

It may be noted that scales significantly correlated with both of the intelligence tests show a similar trend. Thirty-four of the 50 correlations had Analogies greater than Vocabulary yielding a chi-square of 6.50, $p < .02$. There was a slight trend in the opposite direction for the insignificant correlations with 12 out of 18 higher on Vocabulary, but the chi-square of 2.00 gives a p value of only .15.

Finally, a general index of the differential association of interest

with the two intelligence tests and with the two sections of the CMT was made by correlating the columns of correlations in Table 1 using Pearson r . These coefficients are given in Table 4.

Both sexes show the D-48 to be more highly correlated with Analogies than with Vocabulary. For males the values are .90 and .73, for the females they are .86 and .71. Even cross-sex correlations show the same kind of pattern. Male D-48 correlates .90 with female Analogies but only .80 with Vocabulary. Counterpart correlations for females are lower in absolute value but show the same pattern, .76 with Analogies and .58 with Vocabulary.

It should be pointed out that the sexes are quite similar in the association of interest scales and intelligence scores. Intercolumnar correlations for males and females are: total CMT, .94; Vocabulary, .92; Analogies, .95; and D-48, .89.

Some insight into the nature of interests associated with non-verbal intelligence may be gained by examining the correlations in Table 1. For both sexes the highest correlation for the D-48 is with the special SVIB scale S-4 measuring low origence/high intellectence. S-4 has been found to identify persons who are efficient, logical, and methodical; they prefer difficult tasks that can be solved by systematic application of rational procedures derived from conceptions and abstractions as well as by following rules and regulations; in temperament they are introversive though not necessarily asocial (Welsh, in press).

TABLE 4

Intercorrelations of Columns of Correlation Coefficients for Interest Scales with Intelligence Test Measures

Test	Male	Columnar Correlation						
		(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1)	Total CMT	99	98	82	94	94	92	67
(2)	Vocabulary		94	73	91	92	87	59
(3)	Analogies			90	95	94	95	76
(4)	D-48				85	80	90	89
Female								
(5)	Total CMT					99	99	79
(6)	Vocabulary						96	71
(7)	Analogies							86
(8)	D-48							

Note.—Decimal points preceding entries have been omitted.

Other leading correlations for males are in order of magnitude: Mathematician, Physicist, Chemist, Psychologist, Engineer, Physician, Math-Science Teacher, and Senior CPA. For females the highest correlations are: Senior CPA, Chemist, Math-Science Teacher, Psychologist, Physician, Army Officer, Psychiatrist, Mathematician, Physicist, and Engineer. Although there are some interesting sex differences to be noted, in general nonverbal intelligence is associated with scientific interests particularly the physical sciences of Group II and some of the biological sciences of Group I. For females, the sole scale of Group VII, Senior CPA, also shows a marked associated interest.

Highest negative correlations with nonverbal intelligence occur for the special SVIB scale S-1, high origence/low intellectence; again this is true for both sexes. Persons high on S-1 tend to be extravertive in temperament and describe themselves by traits such as adventurous, easy-going, pleasure-seeking, and talkative. They prefer social situations that are not demanding intellectually although they enjoy pitting their wits against others.

Regular SVIB scales showing high negative correlations for males include: Life Insurance Salesman, Real Estate Salesman, Mortician, and Sales Manager. For females the order is: Real Estate Salesman, Life Insurance Salesman, Banker, Mortician, and Sales Manager. This cluster of negative *rs* includes all three sales or business contact scales of Group IX as well as some of the business detail scales of Group VIII.

Vocational interests associated with verbal intelligence may be seen in particular by the correlations of the Vocabulary section of the CMT. Highest position correlations for males are: Psychologist, S-2, Specialization Level, and Psychiatrist. For females they are: S-2, Psychologist, Psychiatrist, and Physician. S-2, the special scale for high origence/high intellectence has been associated with creativity in adults and rated originality in adolescents. A study of poets showed them to score in this section of the two-dimensional model mentioned above (King, 1969).

Subjects high on S-2 describe themselves as complicated, dis-orderly, original, and unconventional. They are introversive (as are S-4's) but in addition are aloof and self-centered. Tasks which are open-ended and unstructured seem to challenge them and they like imaginative solutions to problems. In regular Strong scales

verbal intelligence seems to be related especially to the independent professions and the biological sciences of Group I.

Highest negative correlations appear in the male column for Mortician, S-3, S-1, Pharmacist, Purchasing Agent, Veterinarian, and Policeman. For females the order is: Mortician, S-3, Banker, S-1, Purchasing Agent, and Pharmacist. Most of these regular SVIB scales fall in Group VIII, business detail.

In addition to S-1 which was previously discussed, another special scale, S-3, low origence/low intellectence, also shows significant negative relation. S-3 subjects see themselves as appreciative, energetic, friendly, and practical. They are extravertive and seem to enjoy working with people in a direct relationship. Routine tasks related to tangible matters engage their interest and they prefer a regular, orderly, and systematic approach to problems.

The relationship of vocational interests to intelligence is obviously complex, but evidence from the performance of gifted adolescents suggests several associated trends.

There seems to be a positive relationship between nonverbal intelligence scores and scientific interests in the physical sciences and other vocations stressing methodical and rational approaches to their problems. Business interests, particularly in sales occupations and in vocations requiring social and personal contact with people, show negative relationship with nonverbal intelligence scores.

The previously reported association of verbal-linguistic interests and verbal intelligence was less clearly demonstrated by the correlational approach of the present study although a tendency toward this kind of a relationship was found in the patterns of different intelligence scores. Interests in the professional, biological sciences show a positive relationship to verbal intelligence scores, while interests in business detail and other occupations characterized by routine and systematic procedures seem to be negatively related.

Some evidence was found to justify using part-scores on the CMT² since the kind of intellectual ability and interest required for good performance on Analogies seems to be related to D-48 scores.

²The CMT *Manual* gives only total-score norms, however, complete norms based on the gifted adolescents of the present study are available for the part-scores as well as CMT total and the D-48 (Welsh, 1969).

Finally, some of these observations may be arrayed on a two-dimensional conceptual model that ties together personality traits, vocational interests, and intellectual performance.

REFERENCES

- Black, J. D. *Preliminary manual, the D-48 test*. Palo Alto, California: Consulting Psychologists Press, 1963.
- King, L. D. Personality and aesthetics: Two studies of poetic communication. Unpublished Master's thesis, University of North Carolina, Chapel Hill, 1969.
- Strong, Jr., E. K. *Manual for the Strong Vocational Interest Blanks for Men and Women, Revised Blanks (Forms M and W)*. Palo Alto, Calif.: Consulting Psychologists Press, 1959.
- Terman, L. M. *Manual, the Concept Mastery Test*. New York: The Psychological Corp., 1956.
- Welsh, G. S. *Preliminary Manual, the Welsh Figure Preference Test (research ed.)*. Palo Alto, Calif.: Consulting Psychologists Press, 1959.
- Welsh, G. S. Verbal interest and intelligence: Comparison of Strong VIB, Terman CMT, and D-48 scores of gifted adolescents. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 349-352.
- Welsh, G. S. *Gifted adolescents: A handbook of test results*. Greensboro, N. C.: Prediction Press, 1969.
- Welsh, G. S. *Personality dimensions of creativity*. San Francisco: Jossey-Bass (in press).

MEASURES OF EGO IDENTITY: A MULTITRAIT MULTIMETHOD VALIDATION¹

FRANK BAKER

Harvard Medical School

ERIKSON examines the growth of the personality in terms of a series of eight successive stages, "predetermined in the human organism's readiness to be aware of, and to interact with a widening social radius" (Erikson, 1959, p. 52). As the developing individual encounters different aspects of the social environment, each step becomes a potential crisis because of the attendant radical change in perspective. The problems of each stage can be solved in one of two polar directions which lead to the development of a series of alternative basic senses or attitudes: (a) trust versus mistrust, (b) autonomy versus shame and doubt, (c) initiative versus guilt, (d) industry versus inferiority, (e) identity versus identity diffusion, (f) intimacy versus isolation, (g) generativity versus self-absorption, and (h) integrity versus disgust and despair. The psychosocial quality of each basic attitude becomes more differentiated, as the ego comes into the possession of a more intensive apparatus, even as society challenges and guides such extensions.

Adolescence is the stage at which Erikson postulates a crisis of identity. The term "ego identity" is used by Erikson to denote certain comprehensive gains which the individual, by the end of adolescence, must have derived from all pre-adult experience in order to be ready for the tasks of adulthood. The alternative to the establish-

¹ This article is based on data collected for a dissertation completed for the degree of Doctor of Philosophy at Northwestern University, 1964. The author wishes to express his appreciation to Donald T. Campbell and Gilbert K. Krulee for their guidance and encouragement. The study was supported in part by U. S. Office of Education Project C-998, Contract 3-20-001, under provisions of Title VII of the National Defense Education Act.

ment of a sense of identity is the development of a sense of identity diffusion. While identity diffusion is temporarily unavoidable in this adolescent period of physical and psychological upheaval, the danger is that there may result a permanent inability to "take hold."

While Erikson asserts that "a sense of identity . . . can be defined and evidence from the presence of a dominant attitude of this kind can be described behavioristically" (Erikson, 1950, p. 63), until recently, there has been little empirical investigation of the sense of ego identity. Perhaps the main explanation is to be found in the lack of clarity in Erikson's own conceptualization of the term. He chooses to use his term identity with a number of connotations.

The present study attempts the difficult task of translating this abstract, global, imprecisely defined concept into concrete operational terms. In an attempt to identify the major components of identity, Erikson's clinical description of identity diffusion and his discussion of the healthy personality were closely examined.

In one of his earliest writings on the topic, Erikson describes the central loss of a feeling of identity among young war veterans:

What had broken down . . . was a sense of identity, a sense of who one is, of knowing where one belongs, of knowing what one wants to do. I must emphasize from the outset that this kind of identity is not the same as that which is called 'personal identity.' It is not an amnesia of one's name or history. It is rather a breakdown of the sense that there is continuity and sameness and meaning to one's life history, (Erikson, 1950, p. 16).

In a later essay, Erikson (1959) defines the sense of ego identity as follows:

The sense of ego identity, then, is the accrued confidence that one's ability to maintain inner sameness and continuity is matched by the sameness and continuity of one's meaning for others, (Erikson, 1959, p. 89).

He goes on to describe its concomitants as "a sense of 'knowing where one is going,' and an inner assuredness of anticipated recognition from those who count" (Erikson, 1959, p. 118).

Four aspects of the sense of ego identity were derived from these and other such statements. In contrast with individuals with a diffuse sense of identity, an individual with a well developed sense of

identity: (a) knows who he is, (b) knows where he is going, (c) perceives himself as having "inner sameness and continuity," and (d) is certain about the way his perception of himself compares to the perceptions which others have of him.

Method

Construction of Scales

For each of the four aspects of identity defined above, an eight item Likert-type scale was designed. Each of the thirty-two declarative statements was written following the assumption that if a person strongly agrees with such statements, it would indicate one extreme of the particular characteristic being tapped, and if he strongly disagrees, it would indicate that he possesses the opposite extreme. In order to minimize the effects of an acquiescent response set, each scale was composed of an equal number of positively worded and negatively worded items. Illustrative examples of the items for each scale are presented below:

Knows who he is

It isn't necessary to be a chameleon and be all things to all people in order to get ahead in life.

What a bore it is, waking up in the morning always the same person.

Knows where he is going

The major decisions a person makes are guided by the plans he has for the future.

Life is chaotic, without direction or meaning.

Perceives himself as having "inner sameness and continuity"

A person doesn't change much once he has started out in the world.

No one is the same person from day to day.

Is certain about the way his perception of himself compares to the perceptions which others have of him

A person can be confident of getting recognition from those who count.

What really matters is what other people think; it is not enough just to be sure of oneself.

Five levels of response were provided for all of the items: "strongly agree," "agree," "undecided," "disagree," and "strongly disagree." For positively worded items, "strongly disagree" was scored 5, "agree," 4, "undecided," 3, "disagree," 2, and "strongly disagree," 1. Scoring of negative items was in the reverse direction. Scores on each variable were arrived at by adding up the reversed scores for the particular items composing each of the eight scales.

Sentence-Completion Instruments

Eight sentence stems for each of the four aspects of identity were written. The stems were specifically designed to elicit responses relevant to expressions of one of the four characteristics of identity.

The thirty-two sentence stems were assembled in an alternation format and labeled "Incomplete Sentence Blanks." This instrument was administered as part of the longer questionnaire containing the scales described above. Illustrative sentence completion stems for each of the four aspects are presented as follows:

Knows who he is

When somebody confuses me with someone else, I . . .

Pretending to be somebody you aren't is . . .

Knows where he is going

The things I want out of life are . . .

In making plans for the future, I . . .

Perceives himself as having "inner sameness and continuity"

The person I was yesterday and the person I am today are . . .

If it were possible to go back in time and see myself as I used to be, I would probably feel that I . . .

Is certain about the way his perception of himself compares to the perceptions others have of him

I am sure that people think of me as . . .

In a comparison of the way I see myself and the way others see me, I think . . .

A scoring-by-example manual was developed for scoring the "Incomplete Sentence Blank" following much the same procedure detailed by Renner, Maher, and Campbell (1962). On the basis of pre-

liminary administrations of the instrument, 185 protocols were obtained and were used as a reservoir from which responses were abstracted which seemed relevant to the descriptions of each of the seven variables. Each response was assigned a weight of 5, 4, 3, 2, or 1, depending upon whether *E* judged the item to be: 5—strongly indicative of the positive pole of the variable, 4—somewhat indicative of the positive pole of the variable, 3—ambiguous or could not be scored, 2—somewhat indicative of the negative pole of the variable, 1—strongly indicative of the negative pole of the variable. Each completion was judged only in terms of characteristics of the variable its stem had been designed to elicit.

This pool of items was assembled into a scoring-by-example manual. The end product was a manual made up of illustrations of scorings of sentence completions for each incomplete sentence stem for the one variable the stem had been designed to tap.

Two raters independently scored samples of 20 protocols drawn from the pool of 705 questionnaires used in this study. To reduce the "halo" effect as much as possible, each item was scored on all 20 questionnaires before the rater proceeded to score the next item. The stack of questionnaires was shuffled between the scoring of each item. After this procedure had been completed, inter-rater reliability was computed for each of the variables by computing the product-moment correlation for raw scores. The correlations obtained were: Knows who he is, +0.95; Knows where he is going, +0.96; "Inner sameness and continuity," +0.96; and "Knows his stimulus value," +0.96.

In the major validity study, the description of which follows, each of the 705 protocols was scored by one rater, with similar precautions against a "halo" effect. The scores assigned to each respondent on the sentence completion were the sums of the raters' scores on those completions for the stems of each aspect of identity.

Subjects

The instruments, previously described, were administered during Orientation Week as part of a longer questionnaire to 715 male freshmen entering Lehigh University. Seven-hundred and five questionnaires were returned sufficiently completed to be used in this study and the analysis of data presented here is based upon this sample of 705 male college students.

*Results and Discussion**Reliability*

Reliability was estimated by computation of Kuder-Richardson (Formula 20) reliability coefficients. The reliabilities for the variables under consideration as measured by the Likert-type scales and sentence completion methods are presented in Table 1.

These reliabilities appear to be quite low. In order to check whether or not they are significantly different from zero, the obtained reliabilities could be turned into an F ratio by the formula: $r_{kr} = 1 - 1/F$. This F is for the main effect of persons tested against the person-times-items interaction. The degrees of freedom are as follows: (Persons - 1) and (Persons - 1) (Items - 1). Since we have 705 persons and eight items, the degrees of freedom would be 704 and 4,928. From an F table, we discover that the F required for $p < .01$ with these df is approximately 1.13, which translates into a reliability of .11 as the minimum $p < .01$ level. The more persons, and the more items, the lower the reliability required for significance. Applying this criterion, the obtained Kuder-Richardson reliabilities are all significantly different from zero.

However, since the magnitudes of these reliability coefficients compare unfavorably with those usually reported, one possible explanation should be mentioned. An explanation lies in the nature of the concepts themselves, the definition of which poses complex multi-dimensional constructs. Since increasing complexity of a trait increases the probability that individuals will exhibit unique patternings of the component elements, it is generally true that the more complex the measure, the lower the reliability estimate will

TABLE 1
*Kuder-Richardson Reliability Coefficients for Scales and
Sentence-Completion Measures*

	Scales	Sentence-Completion Measures
Knows who he is	.32	.28
Knows where he is going	.39	.55
Inner sameness and continuity	.48	.31
Knows his stimulus value	.23	.33

be. It is important to note that the Kuder-Richardson reliability coefficient for the *F* scale, which was also administered to this student group at the same time, is only .39 and using that classic instrument as a standard of comparison these new measures compare very favorably.

It can be further argued that internal consistency is irrelevant to the testing of the hypotheses. These items represent an effort at translating some global, abstract concepts into concrete operational terms. The combination of the items is inevitably better at representing the construct in question than any one of them would be. Although in each item some of the specific details introduce tangential values, these irrelevancies differ from item to item, tending to be outweighed, in the total score, by the common core. The item set, whether or not it were to turn out that the individual differences within a measure cohere as a psychological syndrome, would remain, subjectively, the most accurate operational representation of the construct in question.

Relations between Aspects of Identity

Table 2 presents the correlation matrix which resulted from intercorrelating the total scores of the four aspects of identity, as measured by the specially constructed Likert-type scales described above with the total scores of sentence-completion measures of these same variables. Such a matrix of intercorrelations resulting when each of several traits or constructs is measured by each of several methods has been referred to as a multitrait multimethod matrix (Campbell and Fiske, 1959).

Campbell and Fiske have noted that: "Insofar as the traits are expected to correlate with each other, the monomethod correlations will be substantial and heteromethod correlations between traits will also be positive" (Campbell and Fiske, 1959, p. 104). Reversing their emphasis on discrimination between measures of different traits, the concern here is with showing the convergence of related traits (or aspects of a construct) across methods. If traits are predicted to be closely interrelated, they should be significantly intercorrelated when measured by the same methods and when measured by different methods.

Three of these four aspects of identity as measured by the same method and as measured by different methods show significant

TABLE 2

*Multitrait Multimethod Matrix Based on Direct-Attitude and Sentence-Completion Measures of Ego Identity**

	Direct-Attitude Method				Sentence-Completion Method			
	<i>A</i> ₁	<i>B</i> ₁	<i>C</i> ₁	<i>D</i> ₁	<i>A</i> ₂	<i>B</i> ₂	<i>C</i> ₂	<i>D</i> ₂
Direct-Attitude Method								
Knows who he is	<i>A</i> ₁	(.32)						
Knows where he is going	<i>B</i> ₁	.39	(.39)					
Inner sameness and continuity	<i>C</i> ₁	.01	.02	(.48)				
Social stimulus value	<i>D</i> ₁	.43	.21	-.06	(.23)			
Sentence-Completion Method								
Knows who he is	<i>A</i> ₂	.23	.19	.10	.24	(.28)		
Knows where he is going	<i>B</i> ₂	.15	.25	.01	.10	.25	(.55)	
Inner sameness and continuity	<i>C</i> ₂	.18	.13	.32	.18	.10	.02	(.31)
Social stimulus value	<i>D</i> ₂	.18	.16	.05	.24	.24	.15	.10 (.33)

* *df* = 703, *r* = .07 is significant at 5% level; *r* = .10 at 1% level.

intercorrelations in all cases. In the single method triangle representing the intercorrelations of the multiple variables measured by the one method of Likert-type scales, "Inner sameness and continuity" is not significantly correlated to "knows who he is," "knows where he is going," or "social stimulus value." Its correlation with the first two of these is near zero and the correlation with the last is negative and almost at the 5 per cent level of significance. In the sentence-completion triangle it is significantly correlated with "knows who he is" and "social stimulus value" but again is near zero in its correlation with "knows where he is going."

While all four of the defined characteristics of identity have monotrait-heteromethod values which are statistically significant from zero and hence evidence convergent validity, only "inner sameness and continuity" has a validity diagonal value higher than values lying in its column and row in the multiple-methods block. This would indicate that "inner sameness and continuity" is more closely related to itself measured by different methods than to the other aspects of identity.

Consistent with the preceding evidence of independence, "inner sameness and continuity" alone meets the Campbell and Fiske

discriminant validity requirement of correlating higher with itself than other traits which happen to employ the same method. "Inner sameness and continuity," contrary to the theoretical prediction, would appear to be a trait unrelated to the other three aspects of identity.

The data support this prediction that the college student respondents studied here know who they are, know where they are going, and are aware of how their perception of themselves compares to the perceptions others have of them. Apparently they do not necessarily see themselves as having "inner sameness and continuity" as it has been empirically defined here. Conversely, respondents who are not sure who they are, are also not sure where they are going, or what their "social stimulus value is"; do not necessarily lack a feeling of inner sameness and continuity.

Conclusions

This study was successful in translating Erikson's concept of a sense of ego identity into useful operational terms and three of the four aspects of the concept were found to be significantly intercorrelated. "Inner sameness and continuity," appears to be unrelated to the other measures. It is difficult to be conclusive about this finding since any study such as this is both a study of the instruments and the theory on the basis of which the instruments are constructed and their interrelationship predicted. "Inner sameness and continuity" as it was operationalized here also may contain an inflexibility that Erikson had not intended in his use of the concept.

The results of this study lend support to the concept of identity as a significant variable, descriptive of variations of self-attitude among late adolescents and related to the earlier development of a sense of trust versus mistrust. Further research on the reliability and validity of the combined three scales of ego identity developed here is called for in order to test the empirical usefulness of the Erikson theoretical structure.

REFERENCES

- Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- Erikson, E. H. Growth and crisis of the "healthy personality." In M. J. E. Senn (Ed.), *Symposium on the healthy personality, Supplement II: Problems of infancy and childhood*. New York: Josiah Macy, Jr. Foundation, 1950.
- Erikson, E. H. Identity and the life cycle: Selected papers. *Psychological Issues*, 1959.
- Renner, K. E., Maher, B. A., and Campbell, D. T. The validity of a method for scoring sentence-completion responses for anxiety, dependency, and hostility. *Journal of Applied Psychology*, 1962, 4, 285-290.

DIMENSIONS OF PSYCHOPATHOLOGY IN MIDDLE CHILDHOOD AS EVALUATED BY THREE SYMPTOM CHECKLISTS

ELISE E. LESSING AND SUSAN W. ZAGORIN¹

Institute for Juvenile Research
Chicago, Illinois

In spite of widespread agreement regarding the desirability of a theory-derived diagnostic classification system for the psychiatric disorders of childhood (Bard, Sidwell, and Wittenbrook, 1955; Rutter, 1965; Freud, 1965; Achenbach, 1966), no personality theory has obtained sufficient general support to form the basis for such a system. Increasing numbers of investigators have, therefore, focused their efforts upon developing a descriptive classification of manifest symptoms defined in behavioral terms with a minimum of interpretation and inference.

As Miller (1967a) has indicated, many variables contribute to the differences across studies in regard to the descriptive classification scheme derived. Mathematical considerations such as the type of factor analytic rotation employed and the number of factors extracted, subject population considerations such as the types of children evaluated, and item considerations such as number of items representing a given problem area can all influence what emerges as the major syndromes in a given study. Only the bipolar division of symptoms into those involving primarily inner discomfort and those involving negatively valued acts against persons or objects in the environment has been demonstrated to have a generality that transcends most variations in subjects, items, and procedures.

¹ The authors would like to express their appreciation to Michael S. Black, who provided statistical consultation and supervised the canonical variate analysis, and to Merton Krause and Ferdinand van der Veen for helpful comments regarding the manuscript.

Thus, Peterson's (1961) Personality Problem and Conduct Problem, Achenbach's (1966) first principal bipolar factor of Internalizing-Externalizing, the second-order factors of Aggression and Inhibition identified by Miller (1967a), and the Rebelliousness and Anxiety factors of Collins, Maxwell, and Cameron (1962) all reflect the same basic dichotomy.

Other factors, proposed as basic syndromes, have shown less stability across studies. Learning Disability or School Failure has been identified by several investigators, each using a different checklist (Collins, et al., 1962; Brewer, 1961; Miller, 1967a). However, this factor has never emerged in the series of studies utilizing the Peterson Problem Checklist (Peterson, 1961; Quay, 1964; Quay and Quay, 1965). On the other hand, the Inadequacy-Immaturity factor derived from Peterson Problem Checklist data by Quay and his collaborators (Quay, 1964; Quay and Quay, 1965) was not identified by Miller or Brewer, though their subjects were child guidance clinic patients among whom such a cluster of symptoms might be expected on a priori grounds.

The practical problem of selecting a symptom checklist for possible future routine use at the Institute made a salient issue of the extent to which differences in item content would affect one's conclusions regarding the nature of the presenting psychopathology. Previous studies provided considerable data regarding factor stability across samples with symptom items held constant (Quay and Quay, 1965; Quay, Morse, and Cutler, 1966), but very little data regarding factor comparability across item samples with subject sample held constant. Therefore, the present study was undertaken with the following specific purposes: (a) to compare the symptom factors obtained when three different sets of items were utilized for behavior ratings of the same sample of children, and (b) to provide additional information regarding the major factors obtainable from the widely used Peterson Problem Checklist, with particular attention being given to the item content of the Inadequacy-Immaturity factor.

Method

Sample

Subjects were 102 children, aged 10 years 0 months through 12 years 11 months, who received a psychiatric examination at the

Institute for Juvenile Research from September, 1963 through June, 1965, after being referred to the Institute for child guidance services. All admissions during this time period were included in the research sample except for children whose psychiatric interview occurred during vacations of the research personnel conducting the project. The 102 children constituting the sample for the present study were part of a group of 110 children who served as the clinic sample for a validation study of the IPAT Children's Personality Questionnaire (Lessing and Smouse, 1967). The original sample was limited to children with IQ scores of 70 or higher. The present sample was further limited to those children whose mothers evaluated them on both the Peterson Problem Checklist and the Wichita Guidance Center Checklist and whose examining psychiatrist coded the IJR Symptom Checklist on the basis of the mother's oral report of the child's symptoms.

Instruments

The Peterson Problem Checklist (Peterson, 1961) and the Wichita Guidance Center Checklist (Engel, 1955; Brewer, 1961) were selected for use since both are item samples from the universe of referral symptoms reported by parents of disturbed children brought to child guidance clinics. However, the two instruments had yielded different factor structures in previous studies.

The Peterson Problem Checklist consists of the 58 most frequently reported symptoms (e.g. "Disobedience," "Daydreaming") among 477 representatively chosen child guidance clinic cases. The mothers in the present study were merely asked to circle the number of each item which described their child. Although Peterson asked mothers to judge whether the problem was "mild" or "severe" in his original large-scale study, the differentiation of severity was dropped in the statistical analysis he performed (Peterson, 1961, p. 205), and thus did not influence the factor structure he obtained.

The Wichita Guidance Center Checklist (Engel, 1955; Brewer, 1961), like the Peterson Problem Checklist (PPC) contains items based on parental descriptions of the referral problems of pre-adolescent child guidance patients. However, while the PPC items were culled from a large sample of case records, the WGCC items were based on detailed narrative descriptions composed by 25 mothers, supplemented by additional items obtained from 25 other clinic

mothers. The 55 items included in the final checklist significantly discriminated between a group of children of PTA mothers and a group of children referred for child guidance clinic services. The WGCC items are in sentence form, such as: "My child does not seem to be learning like he (she) should" and "My child is a discipline problem, at home and in school." Brewer's (1967) factor analysis of ratings on this checklist made by the mothers of 200 boys, aged 5-13 years, who were referred for child guidance clinic services, yielded five factors: Conflict with Parents, Conflict with Teachers, Failure in Peer Relations, Inner Tension, and School Failure.

The IJR Symptom Checklist (IJRC) is a list of 36 symptoms such as "Stealing," "Shy, withdrawn, timid" etc. which were compiled by the Institute's staff psychiatrists. These items appear on case data summary cards (see Lessing and Schilling, 1966) and are routinely coded by the examining psychiatrist shortly after the diagnostic psychiatric interviews with the mother and child. Like the PPC and the WGCC, the IJRC codifies data regarding the referral symptoms of the child in dichotomous form.

Procedure

The mothers were asked to fill out the two symptom checklists while the child was being interviewed by the psychiatrists. In half of the sample, the PPC was administered first and in the other half the WGCC was given first. The second checklist was introduced with the explanation that the clinic was interested in getting as complete a picture of the child's symptoms as possible and the first list did not cover everything. The mothers were asked to check everything that applied on the second checklist even if it had already been checked on the first. The mother was then interviewed by the psychiatrist, who inquired about the child's problems. The mother's responses were then summarized by the psychiatrist in his coding of the IJR Symptom Checklist.

Statistical Analysis

All checklist items were coded as either 1 or 0. Then a product moment correlation matrix was computed among all items comprising each of the three checklists. Separate principal axis factor analyses were performed upon each of the three correlation matrices. The squared multiple correlation of each variable with all other

variables was used as the communality estimate in the diagonals. The decision regarding the number of factors to be rotated was guided by two initial criteria: (1) retained factors should have an eigenvalue of at least 2.00, and (2) retained factors should account for at least 5 per cent of the variance. However, three factor solutions were tried for each checklist: a solution involving the number of rotated factors dictated by the initial criteria, a solution involving one less than this number of factors, and a solution involving one more than this number of factors. The final choice for each checklist was that factor solution which produced the strongest most distinctive factors (i.e. those with many loadings above .50, but whose defining items have high loadings only on the factor being defined).

The similarity between the PPC factors obtained in this study and those obtained in previous studies was evaluated by means of Tucker's coefficient of factor congruence (Harman, 1960, p. 257). The coefficient consists of the summed products of the rotated factor loadings of all items on the pair of rotated factors being compared divided by the square root of the summed cross products of the factor loadings. The coefficient is interpretable as the equivalent of a correlation coefficient.

In order to evaluate the similarity between the factor spaces defined for each of the three checklists administered in the present study, factor scores on each of the rotated factors were first computed by the method of ideal variables (Harman, 1960, pp. 360-361), and a congruence matrix of factor scores was constructed. The elements of the matrix were coefficients of congruence (Harman, 1960, p. 260) computed between every possible pair of factors. Then a canonical analysis of two checklists at a time was performed in order to find, for each pair of checklists, the linear combination of factors within each checklist which would maximize the congruence between the two factor spaces. That is, for each of a given pair of checklists being compared, a new set of variates (canonical variates) was derived by combining the original factors in such a way that the new variates would correlate maximally with the comparable variates on the other checklist. These canonical variates can be considered as factors resulting from the simultaneous "factoring" of two checklists. The variables "factored" in this case are the binormamin-rotated factors obtained in the initial independent

factoring of each checklist. A canonical variate pattern was obtained for each of the two checklists by postmultiplying the matrix of congruence coefficients among the initial factors by the matrix of canonical coefficients (the weights for combining the factors into canonical variates). This canonical variate pattern is analogous to, and can be interpreted as, a factor pattern. That is, it contains the loadings of the initial factors on the canonical variates. Since the canonical variates are orthogonal, these loadings are the correlations between the initial factors and the canonical variates. As part of the canonical analysis of each pair of checklists, the canonical correlations between the new variates were obtained. The magnitude of these correlations was then compared with that of the congruence coefficients between factor scores on the original, uncombined factors. Direct comparisons were possible because the coefficients of congruence were computed on deviation scores, with a mean of zero, so that they were in actuality correlation coefficients.

Results

Factor Analyses of Three Checklists

The 102 child guidance clinic children comprising the experimental sample showed a high degree of heterogeneity in associations between symptoms. In the case of each of the three symptom checklists, the three or four factors rotated originally accounted for less than half of the total common variance. Rotation of many additional, inefficient factors, accounting for small proportions of the total variance, would have been required to bring the cumulative variance accounted for up to a figure much over 50 per cent.

In the case of the Peterson Problem Checklist (PPC), the four factors rotated accounted for 41 per cent of the common factor variance. A five-factor solution would have met the two criteria that all factors should have eigenvalues of at least 2.00 and account for at least 5 per cent of the variance. However, this solution was rejected because the fifth factor was a weak one with the highest factor loading being .52. Table 1 presents the rotated factor loadings for the four-factor solution along with loadings for the same variables from four other samples of comparable age who were administered the same items. Table 2 contains the values of Tucker's coefficient of factor congruence (Harman, 1960, p. 257) for the PPC factors obtained in the present and the four previous studies.

TABLE 1
*Rotated Factor Loadings for PPC Items with Loadings from Previous Studies of Children
 of Similar Age Presented for Comparison*

Variables	Present Study				Peterson's Fifth and Sixth Grade		Quay and Quay Seventh Grade		Quay and Quay Eighth Grade			Quay's Pre-Adolescent Delinquents		
	I	II	III	IV	I	II	I	II	I	II	III	I	II	III
I. Conduct Problem	68	-18	-09	04	86	11	58	01	62	00	22	64	-03	06
38. Disobedience	64	02	13	01	65	00						51	06	18
47. Destructiveness	59	-08	47	-22	64	16						52	22	-05
27. Temper tantrums	57	04	10	-06	77	07						55	02	-07
25. Fighting	56	-14	-04	16	68	-09	60	07	60	-03	15	69	00	08
11. Boisterousness	56	01	25	-07	69	07						65	16	-15
55. Irritability	55	07	-23	23	76	11	64	14	70	-12	21	67	00	15
8. Disruptiveness	54	-23	-15	19	76	02	48	22	61	-08	19	49	01	-09
3. Attention-seeking	53	11	17	-11	76	08	47	13	33	07	11	57	-16	02
49. Impertinence	51	18	14	-07	60	00						55	06	-03
52. Profane language	50	31	09	-06	71	21	64	21	36	-07	25	59	-04	13
40. Uncooperativeness	50	26	-05	-13	69	28			46	-08	58	50	22	52
22. Inattentiveness	49	-06	08	09	70	15	67	38	29	00	51	53	04	27
48. Negativism	44	37	-09	12	65	20								
33. Irresponsibility	39	25	-30	17	41	13						31	17	40
16. Dislike for school	31	00	27	09	56	11						56	23	-02
17. Jealousy	24	15	-18	11	72	26	62	09	59	-11	36	46	03	53
46. Distractibility	-18	13	07	12	02	16						18	12	16
44. Stuttering														
53. Preference for older playmates	-06	04	00	01	16	01						03	00	-03

TABLE 1—Continued

Variables	Present Study				Peterson's Fifth and Sixth Grade		Quay and Quay Seventh Grade		Quay and Quay Eighth Grade			Quay's Pre-Adolescent Delinquents		
	I	II	III	IV	I	II	I	II	I	II	III	I	II	III
12. Crying	16	20	45	-13	59	19						24	36	17
31. Laziness in school	30	42	-43	11	37	31	57	33	22	-15	59	24	09	66
9. Feelings of inferiority	02	36	38	-03	17	62	01	66	-07	57	17	06	65	05
19. Preference for younger playmates	09	05	34	-15	14	32						17	36	18
42. Passivity	-05	23	30	-07	52	30	47	12	13	14	38	12	28	27
14. Shyness	-19	23	28	-09	-13	51	-35	42	-42	38	00	-27	59	08
56. Enuresis	-15	-08	-23	22								10	08	08
1. Thumb-sucking	03	-02	23	04	05	15								
18. Soiling	-01	-13	14	01										
4. Skin allergy	-13	-12	14	-03	-05	-20								
29. Truancy	07	-08	11	-02	22	35						04	-03	05
IV. Organic-Somatic Problem														
2. Restlessness	20	01	-12	62	71	20	63	12	70	-04	15	44	03	04
37. Tension	17	08	24	56	39	41						57	28	09
26. Nausea	-28	-19	00	52	-02	37								
58. Specific fears	00	23	-01	49	-04	20						08	19	09
54. Nervousness	11	01	19	47	50	26	26	11	28	35	-14	39	40	17
10. Dizziness	-23	-02	14	47										
7. Headaches	-23	03	37	41	00	27								
35. Masturbation	18	-12	33	-34	04	17						-01	27	67
45. Hyperactivity	32	-16	-06	33	49	03	48	-16	54	08	-01	60	-05	09
57. Stomach aches	-21	10	19	27	-06	29								

Note.—Decimals have been omitted to conserve space. In order to facilitate inter-study comparisons, factors obtained in previous studies have all been placed in the same order: Factor I is always Conduct Problem or Psychopathology; II is Prematurity Problems or Neuroticism; and III is Inadequacy-Immaturity or Autism. Where no loadings appear, the particular variable was not analyzed in that sample.

Factor I in the present study has its highest loadings on "disobedience," "difficulty in disciplinary control," "destructiveness," and "fighting." Tucker coefficients of .84, .85, .85, and .88 were obtained when this clear-cut Conduct Problem factor was compared with the Conduct Problem factor obtained in the earlier studies of fifth- and sixth-graders (Peterson, 1961), seventh-graders (Quay, 1964), eighth-graders (Quay, 1964), and pre-adolescent delinquents (Quay, 1966).

Factor II is apparently a combined Personality Problem and Autism factor. The highest loadings are on "preoccupation," "lack of interest in the environment," "excessive daydreaming," "sluggishness," and "lack of confidence." As is indicated in Table 2, Tucker coefficients of .80, .88, .56, and .73 were obtained when this factor was compared with the Personality Problem factor obtained in previous studies. Tucker coefficients of .74 and .73 were obtained when this factor was compared with the Inadequate-Immature and/or Autism factor of previous studies.

Factor III of the current study, with its highest loadings on "self-consciousness" and "hypersensitive," is apparently a Personality Problem variant. Tucker coefficients of .60, .51, .81, and .74 were obtained when it was compared with Personality Problem factors obtained in previous studies.

TABLE 2

Coefficients of Factor Congruence between PPC Factors Obtained in the Present Study and Previous Studies of Children of Similar Age

Present Study	Peterson's Fifth and Sixth Grades		Quay and Quay's Seventh Grade		Quay and Quay's Eighth Grade			Quay's Pre-Adolescent Delinquents		
	I	II	I	II	I	II	III	I	II	III
I Conduct Problem	.84	.12	.85	.16	.85	-.15	.51	.88	.08	.27
II Personality Problem-Autism	.39	.80	.23	.88	-.04	.56	.74	.09	.71	.73
III Personality Problem	.22	.60	-.35	.51	-.33	.81	-.10	.10	.74	.12
IV Organic Problem	.27	.25	.56	.08	.65	-.01	.18	.38	.09	.17

Note.—In order to facilitate inter-study comparisons, factors obtained in previous studies have all been placed in the same order: Factor I is always Conduct Problem or Psychopathic; II is Personality Problem or Neurotic; and III is Inadequacy-Immaturity or Autism.

Factor IV, with its highest loadings on "restlessness," "nausea," "tension," "nervousness," and "specific fears," appears to be an Organic-Somatic factor which has not been found in previous studies utilizing the PPC. The Tucker coefficients of .56 and .65 for the similarity between this factor and the Conduct Problem factor found in Quay and Quay's (1965) samples of seventh- and eighth-graders are not particularly meaningful, since Quay and Quay excluded from their analyses the somatic symptoms having the highest loadings on Factor IV of the present study. Much lower coefficients of factor similarity (.27 and .38) were obtained when Factor IV was compared with the Conduct Problem factors derived when the major portion of the PPC items were used as in the studies of fifth- and sixth-graders (Peterson, 1961) and pre-adolescent delinquents (Quay, 1964).

In the case of the Wichita Guidance Center Checklist (WGCC), four factors, accounting for 48 per cent of the common factor variance, were rotated. Though only three principal axis factors meet the twofold criterion of having eigenvalues of at least 2.00 and accounting for at least 5 per cent of the total variance, the four-factor solution was chosen after inspection of the rotated factor patterns. The four-factor solution produced a strong fourth factor, with several loadings above .50, without disturbing the three previously extracted factors. Table 3 presents the rotated factor loadings for all items with loadings exceeding + or -.35 along with loadings for the same items from Brewer's (1967) factor analysis of data obtained from 200 child guidance clinic boys, aged 5-13 years.

The first factor obtained in the present study has highest loadings on items stating that the child "does not seem to be learning like he should," "makes only passing grades," and "never finishes assignments in school." This cluster is readily identifiable as the School Failure factor obtained by Brewer, and yielded a Tucker coefficient of .86 when compared statistically with the Brewer factor. Factor II has highest loadings on items stating that the child "often does things to attract attention even though he will be punished," "is a discipline problem at home and in school," "frequently gets into things that he knows he shouldn't," is often "hitting and pushing other children," and "is driving his teacher mad." This clear-cut Conduct Problem factor resembles the Conflict with Teacher factor previously obtained by Brewer, and yielded a Tucker

TABLE 3

Rotated Factor Loadings for WGCC Items with Loadings from a Previous Study Presented for Comparison

Variables	Present Study				Brewer Study			
	I	II	III	IV	I	II	III	IV
I. School Failure								
37. Makes only passing grades	68	-18	-23	00	-01	76	-09	10
38. Never finishes school assignments	67	26	-06	-15	09	82	-21	20
5. Not learning like he should	66	-04	06	08	23	74	-07	16
27. Can't keep up with other children	62	-21	24	-18	05	72	34	-05
6. Cannot conform to school tasks	59	23	-08	-06	23	62	-06	31
39. Not ready to do the work expected	56	-11	10	-03	39	60	03	-02
40. Discouraged when having to do something on his own	52	06	19	-23	43	39	39	-31
42. Will not respond in class	52	25	03	-18	-02	66	-11	08
15. Can't get interested in anything	48	-19	27	11	08	44	23	13
25. Very poor reader	48	-07	14	-17	05	74	-04	-07
49. Short attention span	47	17	15	-01	29	29	33	31
1. Seldom finishes things	46	11	00	20	29	38	-04	15
3. Learning under force at home	45	01	12	08	29	19	14	14
20. Lacks self-confidence	45	-13	22	-05	-32	36	49	-15
50. Can't do anything right	42	-06	09	18	-02	20	43	48
44. Has the ability, but won't use it	38	29	-28	14	19	48	-03	38
32. Daydreams a great deal	35	-31	17	02	-14	57	20	01
II. Conduct Problem								
52. Does things to attract attention	06	77	01	-13	39	-18	22	72
51. Frequently gets into things he should not	-03	64	-11	15	66	-06	-02	40
11. Discipline problem (home and school)	01	62	11	11	52	-16	13	63
33. Is driving teacher mad	19	62	-14	-12	-01	00	03	78
9. Hits other children	-03	57	06	-03	-07	-24	14	55
41. Temper-tantrums	-06	53	30	02	56	-24	29	11
55. Restless in school	24	53	-05	-07	-06	-05	42	72
24. Behavior is unpredictable	00	48	24	-17	29	-01	27	52

TABLE 3—Continued

Variables	Present Study				Brewer Study				
	I	II	III	IV	I	II	III	IV	V
54. Jumps or moves around all the time	05	48	16	-12	12	-17	39	56	00
19. Rebellious and resentful	-23	47	36	14	64	-20	26	30	42
48. Has to have everything his way	-08	41	22	22	49	-15	31	33	33
53. Teases and torments other children	06	40	07	05	09	-09	16	49	58
47. Driven to talk constantly	11	36	04	-06	47	-31	37	53	-24
II. Unhappiness-Unsociability									
4. Never seems happy	13	-03	62	-04	16	14	37	01	71
26. Seems to hate everybody	-03	06	54	-08	40	-10	44	-15	50
29. Out of step with family's way of life	17	-03	52	25	29	-09	13	32	32
12. Constantly irritable with children	-12	35	50	-09	12	-26	23	35	69
22. Unhappy all the time	14	-08	46	19	32	17	34	-09	55
14. Can't make friends in school	27	-21	43	-09	-04	00	10	31	75
31. Irritable at home	-14	16	41	19	40	00	52	12	32
IV. Conflict with Parents									
17. Refuses to pick up belongings	-08	-06	-07	79	68	08	-24	11	24
10. Can't keep track of belongings	18	-08	-02	61	18	13	-07	33	11
18. Refuses to help around house	-02	-04	28	58	81	-09	11	18	31
34. No regard for warnings	02	38	-10	47	51	-09	16	56	27
23. Can't get along with father	-02	-12	23	35	28	12	32	-01	51

Note.—In order to conserve space, all decimals have been omitted and all WGCC items have been abbreviated.

coefficient of .78 when compared with the Brewer factor. Factor III has its highest loadings on items stating that the child "never seems happy," "is constantly irritable with the children he plays with," "seems to hate everyone who comes near him," and "is out of step with the way of life in our home." This Unhappiness-Unsociability factor is fairly similar to the Failure in Peer Relations factor identified by Brewer, and yielded a Tucker coefficient of .65 when the two were compared. Factor IV has highest loadings on items stating that the child "refuses to pick up clothes and toys around

the house," and "refuses to do things to help around the house." This factor is moderately similar to the Conflict with Parents factor obtained by Brewer, and yielded a Tucker coefficient of .56 when compared with the Brewer factor. The congruence between the two factors was lowered by the fact that several of the rebellious behavior items which loaded on Brewer's Conflict with Parents factor also loaded on Factor II, Conduct Problem, in the present study. The four-factor solution utilized in the present study did not yield a factor comparable to Brewer's Inner Tension factor.

In the case of the IJR Checklist (IJRC), the three factors rotated accounted for 43 per cent of the common factor variance. Table 4 presents the rotated factor loadings for all items with loadings exceeding + or - .35. Factor I, with its highest loadings on "excessive demands for attention," "disobedient," "immaturity," and "aggressive bullying," may be interpreted as an Immature Conduct Problem factor. Factor II had highest loadings on "truancy from school," "school phobia," and "suicidal attempts," and may be interpreted as a Phobic-Suicidal factor. Factor III, with its highest loadings on "destructiveness" and "stealing," is clearly a second variant of the Conduct Problem factor, and will be labelled Delinquent Conduct Problem.

Comparison of Factor Structure of Three Checklists

Table 5 presents the congruence matrix obtained when congruence coefficients were computed from the factor scores of the 102 experimental subjects on the four PPC factors, the four WGCC factors, and the three IJRC factors. The coefficients representing relationships between factors within tests indicate that even though oblique rotations were performed, the factors within tests are virtually uncorrelated. (The WGCC Conduct Problem and Conflict with Parents is the major exception.) It would appear that the natural simple structure of the checklists is generally orthogonal. The only moderately high coefficient representing congruence across checklists is the .79 obtained between PPC Conduct Problem and WGCC Conduct Problem.

The canonical variate analysis of two checklists at a time revealed that there were linear combinations of factors within checklists which would yield higher congruence coefficients across checklists than the original factors did. The canonical variate pattern for

TABLE 4

Rotated Factor Loadings for IJRC Items

Variables	IJRC Factors		
	I	II	III
I. Immature Conduct Problem			
18. Excessive demands for attention	.64	.07	-.10
19. Disobedient	.58	.18	.07
23. Immaturity	.48	-.22	-.03
15. Poor peer or sibling relationships (aggressive, bullying)	.45	-.11	.37
29. General unhappiness	.40	.00	-.27
17. Defiant and rebellious	.36	.18	.35
II. Phobic-Suicidal			
2. Truancy from school	-.11	.75	.16
14. School phobia	.11	.64	-.05
10. Suicidal attempts	.11	.62	-.12
9. Suicidal threats	.25	.42	-.27
3. Absence from home without parents' knowledge or permission	-.23	.35	.34
III. Delinquent Conduct			
1. Stealing	-.01	.01	.52
7. Destructiveness (other than firesetting)	.08	-.09	.52
16. Poor peer or sibling relationships (passive, victimized)	.11	-.12	-.42
28. Shy, withdrawn, timid	-.06	-.14	-.41

the PPC and the WGCC is presented in Table 6. The first canonical variate for the PPC is almost totally defined by the original Conduct Problem factor. However, the original Organic-Somatic factor also loads upon this first canonical variate, probably because the Organic-Somatic factor includes items such as "restlessness" and "tension," which would be consistent with the acting out focus of the Conduct Problem factor. The first canonical variate for the WGCC is likewise almost totally defined by the original Conduct Problem factor. However, the original Conflict with Parents and School Failure factors also have moderate loadings on this canonical variate. The canonical correlation between factor scores on the new combined factors (canonical variates) is .82 ($p < .01$). This correlation represents no appreciable improvement upon the r of .79 obtained between the original Conduct Problem factors of the two checklists (see Table 5) since the first canonical variate for each checklist in essence coincides with the original Conduct Problem

TABLE 5
Coefficients of Congruence between PPC, WGCC, and IJRC Related Factors

Factors	PPC				WGCC				IJRC		
	I	II	III	IV	I	II	III	IV	I	II	III
PPC											
I Conduct Problem	1.00										
II Personality Problem-Autism	.03	1.00									
III Personality Problem	-.07	-.11	1.00								
IV Organic-Somatic Problem	.11	-.11	.22	1.00							
WGCC											
I School Failure	.23	.45	.04	.23	1.00						
II Conduct Problem	.79	-.04	-.06	.27	.23	1.00					
III Unhappiness-Unsociability	.21	.16	.33	.03	.02	.19	1.00				
IV Conflict with Parents	.35	.14	.06	-.02	.19	.35	.14	1.00			
IJRC											
I Immature Conduct Problem	.29	.01	.06	.07	.09	.20	.14	.06	1.00		
II Phobic-Suicidal	.00	-.15	.04	-.08	-.18	.06	.06	.14	.14	1.00	
III Delinquent Conduct	.48	-.15	-.10	.10	-.04	.51	-.02	.35	-.04	.06	1.00

TABLE 6

Canonical Variate Pattern for Comparison of PPC and WGCC

Original Factor		Loadings on First Canonical Variate	Loadings on Second Canonical Variate
PPC			
I	Conduct Problem	.97	-.06
II	Personality Problem-Autism	.04	.84
III	Personality Problem	-.04	.44
IV	Organic-Somatic	.35	.11
WGCC			
I	School Failure	.35	.74
II	Conduct Problem	.99	-.15
III	Unhappy-Unsocial	.25	.51
IV	Conflict with Parents	.40	.24

Note.—The loadings of each original factor on a canonical variate are the simple correlations between the factor and the canonical variate.

factor. The second canonical variate of the PPC combines the two original factors which contained items involving intra-psychic distress (Personality-Autism and Personality Problem), and can be interpreted as a Personality Problem factor. The second canonical variate of the WGCC is also mainly a combination of the two original factors involving inner discomfort. The canonical correlation between the canonical variates is .60 ($p < .01$). When the original factors were correlated, the highest correlation involving Personality Problem items was .45 between PPC Personality Problem-Autism versus WGCC School Failure.

The canonical variate analysis of the PPC and the IJRC yielded only one significant canonical correlation of .59 ($p < .01$), which did, however, exceed the highest coefficient of congruence (.48) between PPC and IJRC original factors (see Table 1). The first canonical variate of the PPC was again defined almost entirely by the original Conduct Problem factor. The loadings of the original PPC factors on the first canonical variate were: .97 (Conduct Problem), -.19 (Personality Problem-Autism), -.10 (Personality Problem), and .22 (Organic-Somatic). The loadings of the original IJRC factors on the first canonical variate were .48 (Immature Conduct Problem), .03 (Phobic-Suicidal), and .83 (Delinquent Conduct).

The canonical variate analysis of the WGCC and the IJRC also yielded only one significant canonical correlation of .62 ($p <$

.01). Again, the factor common to the two checklists was Conduct Problem. The loadings of the original WGCC factors on the first canonical variate were $-.06$ (School Failure), $.88$ (Conduct Problem), $.04$ (Unhappiness-Unsociability), and $.61$ (Conflict with Parents). The loadings of the original IJRC factors on the first canonical variate were $.23$ (Immature Conduct Problem), $.25$ (Phobic-Suicidal), and $.95$ (Delinquent Conduct).

Discussion

The first purpose of the present study was to investigate factor comparability across item samples with subject sample held constant. When the results of the separate factor analyses and the canonical variate analyses are considered together, the findings suggest only moderate generality of factors across checklists.

Previous research established a high degree of replicability for at least one basic dimension of childhood psychopathology, namely, the internalizing versus externalizing or neurotic versus psychopathic distinction (Peterson, 1961; Collins, Maxwell, and Cameron, 1962; Jenkins, 1964; Quay, 1964; Achenbach, 1966). The present findings are consistent with this earlier work. The canonical variate analysis revealed that a Conduct Problem factor was common to all three checklists, while a Personality Problem factor was common to the PPC and the WGCC.

However, systematic variations in item content across the three checklists resulted in three descriptions of the basic psychopathology of the sample which varied more than many descriptions of different samples evaluated on a single set of items (e.g., see Quay, 1966). The IJRC, for example, did not even yield the classic Personality Problem factor, probably because it contains only two relevant items: "shy, withdrawn, timid," and "general unhappiness." It is also possible that the psychiatrists who rated the IJRC on the basis of mothers' reports were more likely to note the most dramatic symptoms, in contrast to the mothers who themselves filled out the PPC and the WGCC. Thus the only IJRC factor reflecting inner distress has highest loadings for the items "truancy from school," "school phobia," and "suicidal attempts."

The WGCC produced a School Failure factor which did not emerge from either of the other two checklists. The WGCC contains seven items relating to school behavior and academic achievement,

while the PPC contains three, and the IJRC only two. The WGCC School Failure factor is similar to Collins, Maxwell, and Cameron's (1962), Timid, School Failure factor, Dreger et al.'s (1964) Intellectual and Scholastic Retardation factor, and Miller's (1967b) Learning Disability factor, all of which were obtained from child guidance clinic samples. During the nine-year period of 1951-1960, 9.5 per cent of 6483 children referred to the Institute for Juvenile Research for child guidance services had a learning problem as the primary problem area (Lessing and Schilling, 1966, p. 325). Primary problem area was not psychiatrically rated for the present sample consisting of patients from the same clinic. However, it is evident that school failure emerges as an important syndrome when there is sufficient item density (as on the WGCC) to permit such a clustering.

The PPC likewise yields an important clinical syndrome that does not emerge from either of the other two checklists. PPC Factor II, which was labelled Personality Problem-Autism, contains a cluster of items such as "sluggishness, lethargy," "preoccupation," "anxiety, chronic fearfulness," and "easily flustered and confused," which delineate a regressive behavior pattern not obtainable from the WGCC or the IJRC.

Some of the checklist-specific factors are of questionable status and would require further replication before being considered as major, stable psychopathological syndromes. For example, PPC Factor IV Organic-Somatic with its combination of items reflecting tension and irritability with items describing physical symptoms may be a variant of either the Hyperactive, Brain-Injured syndrome of Jenkins and Glickman (1946), or the Somatic Complaints factor of Achenbach (1966, p. 18). WGCC Factor IV, Conflict with Parents, is probably a variant of Factor II, Conduct Problem. In fact, the correlation of .35 ($p < .01$) between WGCC factors II and IV was the one exception to the general orthogonality of the intra-checklist factors (see Table 5).

The second purpose of the study was to provide additional information regarding the major factors obtainable from the PPC, with particular attention being given to the item content of the Inadequate-Immature or Pre-Psychotic factor. The Conduct Problem factor is extremely stable even when it is derived from varying subsets of the PPC items. Previous investigators eliminated in-

frequently endorsed items, with 10 per cent generally being the minimum acceptable percentage of endorsement (Quay and Quay, 1965; Quay, 1966). Thus the number of items actually subjected to factor analysis varied from 26 for the seventh-grade sample studied by Quay and Quay (1965) to 55 for the fifth- and sixth-graders studied by Peterson (1961). All 58 items were used in the factor analyses conducted for the present study. Yet the Tucker coefficients reported in Table 2 all exceed .80 when the Conduct Problem factor is compared across studies.

The clustering of items representing internalized symptoms shows much less stability. Historically, the Personality Problem factor was identified first (Peterson, 1961). The next study utilizing the PPC (Peterson et al., 1961) produced two factors upon which the internalized items loaded heavily. Thus "sluggishness, lethargy" and "preoccupation," which had been among the 15 symptoms interpreted as defining the Personality Problem factor in the original study, now loaded upon the new third factor, labelled "Autism." In subsequent studies, the third factor, which was re-labelled "Inadequacy-Immaturity," proved to be rather unstable in item content (Quay and Quay, 1965; Quay, 1966). In fact, Quay and Quay (1965, p. 218) reported a Tucker coefficient of .44 for the congruence between the Inadequacy-Immaturity factor obtained in their eighth-grade sample and the same factor obtained from Quay's adolescent delinquent sample. However, a Tucker coefficient of .67 was reported for the congruence between the eighth-grade Inadequacy-Immaturity factor and the seventh-grade Personality Problem factor. In the present study, Factor II was labelled Personality Problem-Autism in order to emphasize the mixture of anxious, worrying, self-depreciating neurotic characteristics with the withdrawn, unresponsive, regressive items. Factor II was found to be almost equally congruent with the Personality Problem and Inadequacy-Immaturity factors obtained in previous studies (see Table 2).

Both the present and previous findings in regard to the internalizing items on the PPC may be a reflection in the area of childhood psychopathology of a phenomenon noted by Eysenck (1955). When four objective tests were administered to 20 normals, 20 neurotics, and 20 psychotics ranging in age from 20-40, two orthogonal canonical variates were obtained and labelled "neuroticism" and "psychoticism." It was found that high psychoticism scores were nearly

always combined with high neuroticism scores, though high neuroticism-low psychoticism was the characteristic pattern for the neurotics in the sample. The validity of this interpretation can be evaluated only by administering the PPC to a sample containing comparable proportions of neurotic and psychotic or pre-psychotic children.

The findings of the present study highlight the need for data covering the full range of psychopathology from an adequate sample of clinical types. The collection of such data must be preceded by careful consideration of the problems involved in defining the universe of symptom items and sampling it adequately. Loevinger (1965) has described the difficulties inherent in conceiving of items, rather than persons, as a population to be sampled. The majority of investigators concerned with basic symptom factors have appeared to be guided by the implicit view that the universe to be sampled consists of all presenting symptoms reported by parents or other observers of disturbed children. However, there has been insufficient attention paid to the consequences of sampling from this hypothetical universe in various ways.

The three methods used to construct the symptom checklists used in the present study (selecting the most frequently reported symptoms of representative clinic cases, selecting the symptoms providing statistically significant discrimination between normal and clinic cases from a pool of items provided by 50 mothers of clinic cases, and selecting subjectively judged important symptoms) produced three quite different item samples. When diagnostic symptom checklists are being constructed, sampling methods should be chosen on the basis of their ability to produce an item sample that is optimal for the major purposes of diagnostic classification: parsimonious description, investigation or etiology, determination of prognosis, selection of treatment method, and evaluation of treatment outcome. All of these purposes require factor generality across subject and item samples to an extent that will permit the accumulation of comparable data and the applied use of research findings in a variety of clinical settings. Therefore, it is necessary to do more than merely avoid obviously deficient sampling methods, such as the haphazard procedure used to compose the IJRC, and questionable methods such as the use of only 50 mothers to generate symptom items for the WGCC. Even the results of a single systematic

sampling of items such as that used in the construction of the PPC may require supplementation with items defining major factors obtained from other samples.

The accumulation of independent studies, each based on different item samples, derived from different subject samples, can be regarded as the equivalent of many samplings from the universe of symptoms of disturbed children. Major factors which are consistently obtained across several independently and systematically obtained sets of symptom items may be considered to constitute the ideal factors for organizing the available pool of symptom data. Any specific item sample or checklist whose item content will not yield one of these replicated factors can be considered to be unrepresentative of the hypothetical universe of symptoms, and thus in need of modification. The construction of new symptom checklists should ideally involve the use of marker variables from previous studies as well as symptom items obtained from the subject sample under immediate consideration. Miller (1967a) provided a noteworthy example of the use of this principle in constructing the Louisville Behavior Check List for Males 6-12 Years of Age.

In accordance with the line of reasoning just presented, it would be advisable to add more school performance items to the PPC and more autism and inner distress items to the WGCC. It would then be possible for each of these checklists to yield the Personality Problem, Conduct Problem, Learning Disability, and Autism factors which have the greatest generality across studies (Peterson, 1961; Collins, Maxwell, and Cameron, 1962; Quay and Quay, 1965; Achenbach, 1966; Miller, 1967a). The IJRC does not provide sufficiently balanced coverage of symptoms to serve as a diagnostic instrument.

Summary

The purposes of the study were: (a) to compare the symptom factors obtained when three different sets of items were utilized for behavior ratings of the same sample of children, and (b) to provide additional information regarding the major factors obtainable from the widely used Peterson Problem Checklist. Subjects were 102 child guidance patients, aged 10 years 0 months through 12 years 11 months. All were rated on the Peterson Problem Checklist (PPC) and the Wichita Guidance Center Checklist (WGCC) by their mothers, and on the Institute for Juvenile Research Checklist (IJRC) coded by the examining psychiatrist on the basis of the mothers'

oral report of their children's symptoms. The PPC yielded four factors: Conduct Problem, Combined Personality Problem-Autism, Personality Problem, and Organic-Somatic Problem. The WGCC yielded four factors: School Failure, Conduct Problem, Unhappiness-Unsociability, and Conflict with Parents. The IJRC yielded three factors: Conduct Problem, Phobic-Suicidal Syndrome, and Delinquent Conduct Problem. Even when the techniques of canonical variate analysis were used to combine factors within checklists in order to maximize congruence across tests, only moderate generality of factors across tests was obtained. A reconstituted Conduct Problem factor was common to all three checklists and a reconstituted Personality Problem factor was common to the PPC and the WGCC.

The results of the study were considered to highlight the need for careful attention to the problem of symptom item sampling in the construction and revision of symptom checklists. It was suggested that symptom checklists with the degree of factor structure comparability required for diagnostic purposes can most readily be constructed by supplementing the results of any single sampling of the universe of symptom items with marker variables defining major factors obtained in other studies.

REFERENCES

- Achenbach, T. M. The classification of children's psychiatric symptoms: A factor analytic study. *Psychological Monographs*, 1966, 80, No. 7.
- Bard, J. A., Sidwell, R. T., and Wittenbrook, J. M. A practical classification for emotionally disturbed children treated in a welfare setting. *Journal of Nervous and Mental Disease*, 1955, 121, 568-572.
- Brewer, J. E. A checklist of child behavior problems for use by parents. Unpublished manuscript, Wichita Guidance Center, Wichita, 1961.
- Brewer, J. E. Factor analysis of Wichita Guidance Center Checklist for parents. Personal Communication, 1967.
- Collins, L. F., Maxwell, A. E., and Cameron, K. A factor analysis of some child psychiatric clinic data. *Journal of Mental Science*, 1962, 108, 274-284.
- Dreger, R. M., Reid, M. P., Lewis, P. M., Overlade, D. C., Rich, T. A., Taffel, C., Miller, K. S., and Flemming, E. L. Behavioral classification project. *Journal of Consulting Psychology*, 1964, 28, 1-13.
- Engel, Mary. The development of a scale to be used in play therapy research. *Transactions of the Kansas Academy of Science*, 1955, 58, 561-565.
- Eysenck, H. J. Psychiatric diagnosis as a psychological and statistical problem. *Psychological Reports*, 1955, 1, 3-17.

- Freud, Anna. *Normality and pathology in childhood*. New York: International Universities Press, 1965.
- Group for the Advancement of Psychiatry. Psychopathological disorders in childhood: Theoretical considerations and a proposed classification. New York: GAP Report No. 62, Vol. 6, June, 1966.
- Harman, H. *Modern factor analysis*. Chicago: University of Chicago Press, 1960.
- Jenkins, R. L. Diagnosis, dynamics, and treatment in child psychiatry. *Psychiatric Research Report 18*, American Psychiatric Association, October, 1964, 91-120.
- Jenkins, R. L. and Glickman, Sylvia. Common syndromes in child psychiatry. *American Journal of Orthopsychiatry*, 1946, 16, 244-261.
- Lessing, Elise E. and Schilling, F. H. Relationship between treatment selection variables and treatment outcome in a child guidance clinic: An application of data-processing methods. *Journal of the American Academy of Child Psychiatry*, 1966, 5, 313-348.
- Lessing, Elise E. and Smouse, A. D. Use of the Children's Personality Questionnaire in differentiating between normal and disturbed children. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1967, 27, 657-669.
- Loevinger, J. Person and population as psychometric concepts. *Psychological Review*, 1965, 2, 143-155.
- Miller, L. C. Louisville Behavior Check List for Males, 6-12 Years of Age. *Psychological Reports*, 1967, 21, 885-896. (a)
- Miller, L. C. Dimensions of psychopathology in middle childhood. *Psychological Reports*, 1967, 21, 897-903. (b)
- Peterson, D. R. Behavior problems of middle childhood. *Journal of Consulting Psychology*, 1961, 25, 205-209.
- Peterson, D. R., Becker, W. C., Shoemaker, D. J., Luria, Z., and Hillmer, L. A. Child behavior problems and parental attitudes. *Child Development*, 1961, 32, 151-162.
- Quay, H. C. Personality dimensions in delinquent males as inferred from the factor analysis of behavior ratings. *Journal of Research in Crime and Delinquency*, 1964, 1, 33-37.
- Quay, H. C. Personality patterns in pre-adolescent delinquent boys. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1966, 26, 99-110.
- Quay, H. C., Morse, W. C. and Cutler, R. L. Personality patterns of pupils in special classes for the emotionally disturbed. *Exceptional Children*, 1966, 32, 297-301.
- Quay, H. C. and Quay, Lorene C. Behavior problems in early adolescence. *Child Development*, 1965, 36, 215-220.
- Rutter, M. Classification and categorization in child psychiatry. *Journal of Child Psychology and Psychiatry*, 1965, 6, 71-83.
- Spivack, G. and Levine, M. The Devereux Child Behavior Rating Scales: A study of symptom behaviors in latency age atypical children. *American Journal of Mental Deficiency*, 1964, 68, 700-717.

INDIVIDUAL DIFFERENCES IN DIAGNOSTIC JUDGMENTS OF PSYCHOSIS AND NEUROSIS FROM THE MMPI¹

NANCY WIGGINS

University of Illinois

THE numerous comparisons of clinicians with computers in forecasting behavior (usually to the computer's advantage) have led investigators to the more fundamental problem of how different clinicians arrive at their predictions; for a recent review see Goldberg (1968). This concern with the clinical judgment *process* has necessarily involved the notion of individual differences among clinical judges. As with most kinds of data, individual differences in clinical judgments can be treated in two ways. Should differences among judges exist, these differences can be treated as error, and an "average" judge would be said to provide the most meaningful summary for all of the judges. This approach provides generalizability to other clinicians similar to the ones being studied, at the expense of ignoring individual differences. Thus, it is assumed that the mean or average judgment is representative of all of the judges in the sample, or, alternatively, that the judges are replications of one another within error of measurement.

However, individual differences become important in the context of studying the clinician's judgmental processes. It was in this spirit that Hoffman (1960) compared the judgmental models of two subjects asked to judge the intelligence of persons represented by a series of nine-cue profiles. Hoffman provided a "paramorphic" model of each judge by obtaining the regression weights from the

¹ This research in individual differences in human judgments was supported in part by NIMH Grant No. 13892 to the present author. The thoughtful and time consuming editorial assistance of Lewis R. Goldberg is gratefully acknowledged.

multiple regression of the nine input cues on the judgments themselves. These weights provided a model indicating the relative emphasis a single judge placed on each cue in this context. The major problem with enumerating and comparing the results of clinical judges, individually, is that of generalizability. Thus, one would be willing to generalize the results, or model, for a single judge *only* to that same judge for future occurrences of a similar task.

A slightly different approach to individual differences in human judgment is found in the theoretical models of Tucker and his associates (e.g., Tucker and Messick, 1963; Tucker, 1960). By specifying conceptual types of judges, this approach allows for individual differences to emerge, while providing considerably more parsimony than is the case when each judge is treated individually. The results, or model for different *types* of judges can be generalized to similar samples of judges; it would not be expected that an "identical twin" of any single judge would emerge in a new sample of judges but only that a similar *type* of judge would be generalizable.

Specifically, Tucker's approach involves the factor analysis of sums of squares and cross-products among judges across judgments. This is directly analogous to an obverse factor analysis of subjects rather than variables. Subject factors are then positioned in such a way as to represent meaningful "conceptual" or "idealized individuals" (Cliff, 1968). Thus, a subject factor (or "theoretical subject") is passed through real subjects in such a way as to represent a subgroup of response-homogeneous judges. Should the number of subject factors be one, it would be difficult to argue the case for individual differences in judgment. On the other hand, should the number of subject factors be greater than one, it would be important to isolate meaningful "idealized" types and to determine the personality correlates of these idealized individuals.

Although a variety of studies using Tucker's model have attested to the importance of individual differences in judgmental viewpoints (Messick and Kogan, 1966; Pederson, 1962; Skager, Schultz, and Klein, 1966; Walters and Jackson, 1966; Wiggins, 1966; Wiggins, Hoffman, and Taber, 1969; Wiggins and Fishbein, 1969; Wiggins and Wiggins, 1969; Snyder and Wiggins, 1970), this model has not been widely applied to the area of

clinical judgment. Using Tucker's idealized type approach, the present study involved an analysis of clinical judgment data originally collected by Meehl (1959). In particular, 29 clinicians were required to make diagnoses of psychosis vs. neurosis on an 11-step, forced-normal distribution for 861 MMPI profiles from seven hospitals and clinics around the country. One study utilizing these data (Horn and Stewart, 1968) approached the problem of possible individual differences among these 29 judges in a manner similar to that taken by the present study. Horn and Stewart factor analyzed the diagnostic judgments of the 29 clinical judges, as well as the actual criterion (hospital diagnosis), across the 861 MMPI profiles and retained three subject factors. A varimax rotation of these three factors indicated that the first factor was also marked by the criterion, suggesting a validity component of judgment. No interpretation of the remaining two factors was made.

This inability to interpret the second and third subject factors is not surprising in view of the fact that few personality or judgmental measures of the clinicians were utilized in the analysis. Except for the variable "amount of clinical training" (which was unrelated to any of the factors), Horn and Stewart had no variables by which to identify their subject factors. This is particularly unfortunate in light of the considerable body of research on these same 29 clinicians (Goldberg, 1965, 1968, 1969, 1970; Wiggins and Hoffman, 1968).

In addition to the lack of personality and judgmental correlates for Horn and Stewart's three subject factors, another aspect of the analysis should be noted. Horn and Stewart performed a varimax rotation of their three subject factors. There is no guarantee that these varimax factors represented meaningful "idealized individuals," i.e., passed through real subjects. A plot of Horn and Stewart's varimax factors revealed that these factors were rather poor representations of real subjects; the factors did not pass through, or near, any of the 29 judges. Thus, the three subject factors did not represent real clinicians, and with the exception of the first factor, the remaining two factors were uninterpretable.

In the present study these data were re-analyzed in light of the foregoing considerations. It was hypothesized that individual differences in judgmental viewpoints would emerge. It was further predicted that such differences would be manifested in significant

personality and judgmental correlates of the idealized subject types, when these idealized individuals represented real judges in the subjects' factor space. Psychometrically, these hypotheses can be restated: (a) More than one subject factor will be necessary to account for the sums of squares and cross-products among clinicians; (b) these factors can be rotated in such a way as to represent meaningful idealized individuals, i.e., subgroups of response-homogeneous judges; and (c) these different idealized individuals will differ on available judgmental and personological measures. At the least, it would be expected that if different idealized individuals emerge, their paramorphic judgmental models should distinguish among them.

Method

Subjects. The data utilized in the present study were originally collected by Meehl (1959) and have been described extensively elsewhere (Goldberg, 1965, 1968, 1969, 1970). Thirteen of the subjects were Ph.D. clinical psychologists (staff) and the remaining 16 subjects were predoctoral trainees at the University of Minnesota. These 29 judges were given seven samples of MMPI profiles, one sample at a time, and were asked to sort each group of profiles on an eleven-step forced-normal distribution ranging from most (likely) neurotic to most (likely) psychotic. Each MMPI profile consisted of eight clinical scales (excluding *Mf*) and three validity scales. The only information given the clinicians was that the samples represented males under psychiatric care who were diagnosed as psychotic or neurotic. In fact, the percentage of psychotics in each sample ranged from 37% to 64%, with a median of 51% over all 861 profiles.

Judgmental and personological variables. From the extensive research on these 29 clinicians, a variety of judgmental and personological variables were available. The judgmental variables² taken primarily from Goldberg (1970) were obtained separately for each of the 29 clinical judges, based on analyses using the total group of 861 MMPI profiles. Goldberg (1970) described these variables as follows:

a. *Validity coefficient of the judge:* The correlation between the judge's predictions and the actual criterion values.

² The data for the present study were obtained from Lewis R. Goldberg.

b. *Linear predictability of the judge*: The multiple correlation between the eleven MMPI scale scores and the judge's predictions.

c. *Reliability of the judge*: Correlations between a judge's responses to 100 pairs of empirically matched profiles.³

d. *Validity of the judge's linear model*: The correlation between the predicted judgments based on the judge's linear regression model and the actual criterion values.

e. *Linear component of judgmental accuracy*: The correlation between the predicted values from the judge's linear regression model and the predicted values for the linear model relating MMPI scale scores to the criterion. This variable is perfectly correlated with the validity of the judge's model (d) and provides an alternative interpretation to that measure.

f. *Nonlinear component of judgmental accuracy*: The correlation between residual values of the criterion and the residuals of the judge's predictions after the linear components are removed.

g. *Incremental validity of model over judge*: The arithmetic difference between the validity of the judge's model and the actual validity of the judge ($d - a$).

h. *Relationship to composite judge*: The correlation between the clinician's judgments and those of the "composite" judge (the average of all of the 29 judgments for each profile).

Most of these judgmental variables stem from Tucker's (1964) and Hammond, Hursch, and Todd's (1964) formulation of clinical judgment in terms of the Brunswick lens model. Further, these variables have been shown to be of considerable importance in Goldberg's (1970) recent work comparing the validity of the judge with the validity of his model.

In addition to the above variables, a few personological and demographic variables were obtained: sex, staff vs. trainee, and ratings of "likeability" and "relative intelligence" by a psychologist who knew all but two of the 29 judges. Moreover, for each judge the standardized regression weights for each of the eleven MMPI scales were obtained by regressing the 11 MMPI scale scores onto the 861 judgments. These weights constituted eleven separate variables for each judge.

³ Goldberg, L. R. Personal communication.

Method of analysis. First, sums of squares and cross-products⁴ were obtained among the 29 judges across the 861 MMPI profiles. This matrix of cross-products was subjected to a principal components analysis; the number of factors retained in this analysis was based on an examination of the successive distribution of eigenvalues, as well as on Tucker's (1966) mean square ratio test for factor significance. Each of the retained unrotated subject factors was correlated with all of the outside judgmental variables in order to identify the unrotated principal components. Next, these principal components were hand rotated to represent meaningful idealized individuals (Cliff, 1968) by passing each vector through a single judge. The projections of each of the 29 judges on the idealized individuals (vectors) were obtained, a procedure which is directly analogous to factor rotation in the space of individuals. Again, the correlation of each idealized individual with all of the judgmental and personological variables was obtained in order to identify the different judgmental points of view.

Results

Three subject factors were extracted from the matrix of inter-subject cross-products. Two criteria were invoked to determine the "significance" of these three subject factors: (a) The distribution of successive eigenvalues indicated a large drop in variance after the third factor, with the remaining successive differences approaching an arbitrarily small constant. The eigenvalues were: 903055, 7020, 3409, 2288, 2191, \dots , 484. The first three eigenvalues accounted for 97 per cent of the sums of squares; (b) Tucker's (1966) mean square ratio test, an approximate F -test for factor significance, indicated significant F -ratios ($p < .05$) for the first three factors. Although the first eigenvalue is large relative to the remaining ones, this is due to the recovery of the means by the

⁴Tucker (1968) has argued that when individuals are factored across a common rating scale, sums of squares and cross-products among subjects should be utilized instead of intercorrelations. Thus, any differential mean effect would be uncovered by the first principal component, thereby maximizing the possibility of individual differences. However, for the present data, the rating scale was a forced-normal distribution and as such the means were identical for all subjects. Although the present analysis utilized cross-products following Tucker and Messick's individual differences model (1963), it is noted that the results would be quite similar to an analysis of intercorrelations.

first principal component when cross-products, rather than correlations, are factored.

Table 1 presents the significant ($p < .05$) correlations between each of the judgmental and personological variables and the three unrotated principal components. The first unrotated principal component was highly correlated with the judges' relationship to the composite judge. This is not surprising since the first component was a general factor which essentially recovers the original means in the analysis. These means, of course, would be related to the composite judge. In addition, the first principal component correlation between this component and the judges' validities ($r = .87$). Component II represented a validity factor, as seen by the high correlation between this component and the judge's validities ($r = .87$). An even higher correlation of this principal component was found with the validity of the judges' models ($r = .94$), indicating that this factor represents essentially valid judges whose models tend to outperform them in predicting the actual criterion. This result can be found in Goldberg's (1970) comparison of judges and their models. Thus, the more valid the judge, as represented by principal component II, the more valid his model. In addition, the judges at one pole of component II tended to be 'liked' and to be rated as relatively more 'intelligent' than those at the other pole. Unrotated component III represented both a lack of linear predictability as well as a lack of reliability. Neither sex nor

TABLE 1
Significant Correlations between Three Unrotated Principal Components and Judgmental Variables

Judgmental Variables	Principal Components		
	I	II	III
1. Validity		.87	— .62
2. Linear predictability	.79		— .57
3. Reliability	.66		
4. Validity of model		.94	
5. Linear component of accuracy		.94	
6. Incremental validity of model over judge		.52	
7. Correlation of judge with composite judge	.80		— .46
8. "Likeability"		.46	
9. "Intelligence"		.39	

Note.—Only correlations $\geq .36$ ($p < .05$) have been tabled.

amount of training was related to the three components, nor was the nonlinear component of accuracy.

Since the first principal component tended to be a general factor with little variability among subject loadings, a plot of the second and third unrotated principal components was examined. This plot is presented in Figure 1. As noted earlier, component II, a bipolar factor, discriminated the most valid judges from the least valid judges, and factor three pulled out a single idiosyncratic judge. Lines were drawn in Figure 1 which connected three judges marking the second and third principal components. The triangle thus formed illustrates the configuration of subjects with respect to the second and third unrotated principal components.

With reference to Figure 1, the three unrotated principal components were rotated in such a way that each factor directly passed through each of the three marker judges (end points of the triangle). The projections of all of the remaining judges on these

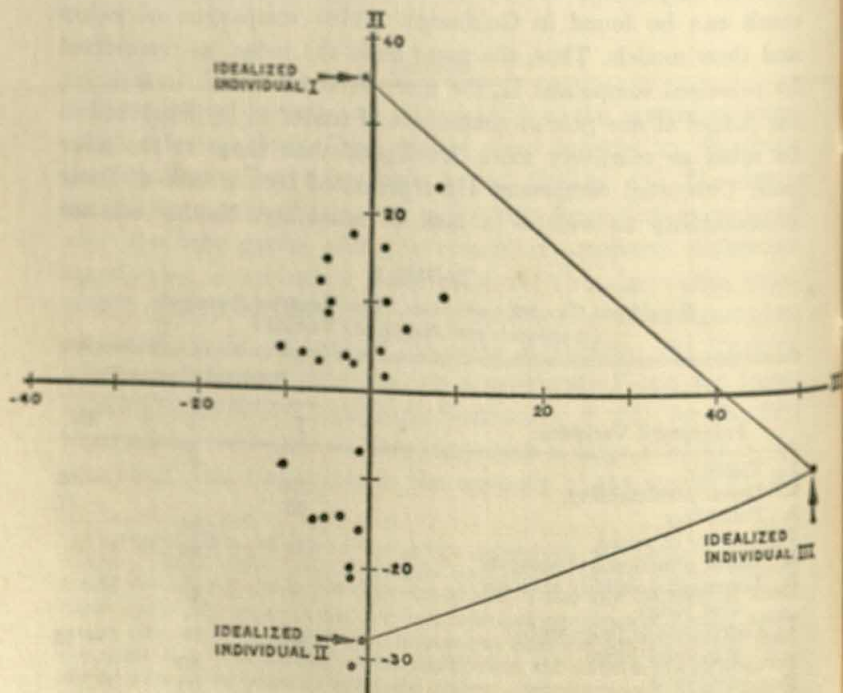


Figure 1. Plot of 29 clinicians on principal components II and III.

three rotated factors were obtained. Each rotated factor thus continuously represented a meaningful "idealized individual;" that is, the factors were relatively unipolar with respect to all of the judges, they were marked by three real judges, and they circumscribed all of the subjects in the subject factor space. In addition, the factor rotation was performed in such a way that each of the three marker judges had a loading of unity on the corresponding rotated factor with zero loadings on the other two factors.³

Table 2 presents the significant correlations ($p < .05$) between

TABLE 2

Significant Correlations between Three Idealized Individuals and Judgmental Variables

Judgmental Variables	Idealized Individuals		
	I	II	III
1. Validity	.93	-.58	
2. Linear predictability		.44	-.62
3. Reliability			-.57
4. Validity of model	.97	-.69	
5. Linear component of accuracy	.97	-.69	
6. Incremental validity of model over judge	.45	-.55	
7. Correlation of judge with composite judge			-.46
8. "Likeability"	.44	-.41	
9. "Intelligence"	.44		

Note.—Only correlations $\geq .36$ ($p < .05$) have been tabbed.

the three idealized individuals and each of the judgmental and personological variables. Idealized individual I, the vector passed through the more valid judges, represented the validity component of judgment ($r = .93$). This idealized individual tended to linearly match his judgments with the linear component of the criterion ($r = .97$). Clinicians represented by this idealized individual tended to be liked and to be rated as relatively more intelligent than the other judges. The significant correlation with the incremental validity of the judge's model indicated that the models for these judges were more valid than their own judgments.

³ It is noted that since the three marker judges were not orthogonal to one another in the subject factor space, the corresponding idealized individuals will be correlated. The intercorrelations between the subjects' projections on the three idealized individuals were: I-II, $r = -.71$; I-III, $r = -.50$; and II-III, $r = -.25$.

Idealized individual II represented judges who were not valid but whose judgments were linearly predictable from a multiple regression equation relating MMPI scales to judgments. The models for these judges were considerably worse in predicting the criterion than their own judgments, both model and man being relatively invalid. This group of judges tended to be disliked. Idealized individual III, marked primarily by a single, idiosyncratic judge, represented an unreliable judge whose judgments were not linearly predictable from the MMPI scales. This idealized individual was also negatively related to the composite judge. Although idealized individual III did not correlate significantly with validity ($r = .34$), the single judge marking idealized individual III was the second most invalid judge with a validity coefficient of .15.

A comparison of these three idealized types of judges indicated that for the valid idealized judge, his model was more valid than his judgments. The reverse was true for the invalid but linearly predictable idealized judge: his model was worse than his judgments. Neither of these two idealized judges was correlated significantly with reliability. In comparison, for the unreliable and less linearly predictable idealized judge, the validity of his model fared about as well as his own judgments. These results support Goldberg's (1970) theoretical comparison between man and his model: when man has any positive validity the model will outperform the man. As man becomes more linearly predictable, the validity of his model approaches the validity of his own judgments.

The variables which did not discriminate among idealized individuals were sex, amount of training, and the nonlinear component of accuracy. Although the present idealized individuals were chosen in such a way as to represent three single judges who marked the second and third principal components, it would be possible to rotate the subject factors in such a way as to maximize the correlations of the idealized individuals with the various judgmental variables. The present positioning of idealized types indicated the major judgmental correlates to be validity, predictability and reliability. Although both idealized individual II and idealized individual III tended not to be valid judges, they differed in that idealized individual III also represented the least reliable judges. It is possible that idealized judge II simply weighted the MMPI

scales incorrectly, whereas the invalidity of idealized judge III stemmed primarily from his unreliability.

Some light is shed on the issue of scale weighting by examination of Table 3. For each judge a multiple regression equation between the eleven MMPI scales and the 861 profile judgments yielded a set of 11 regression weights. These weights reflect the manner in which each judge used the 11 MMPI scales in arriving at his prediction of the criterion. These standardized regression weights for each scale were correlated with the judges' projections on the idealized individuals. The resulting significant ($p < .05$) correlations between regression weights and judges' loadings on the idealized individuals are presented in the left side of Table 3. These correlations indicate the relative importance or the variability a given MMPI scale has for a given idealized individual.

TABLE 3

A Comparison of Idealized Individuals and Marker Judges on Regression Weights Attached to the Eleven MMPI Clinical Scales

Clinical Scales	Significant Correlations Idealized Individual			Standardized Regression Weights			
				Criterion	Marker Judges		
	I	II	III		I	II	III
<i>L</i>				.12	.03	.00	.02
<i>F</i>		.42		-.02	.01	.19	.00
<i>K</i>				.06	.00	-.11	-.01
<i>Hs</i>	-.39	.68	-.47	-.04	-.16	.04	-.24
<i>D</i>	-.60	.62		-.05	-.29	.13	-.09
<i>Hy</i>	-.51		.47	-.28	-.27	-.07	.01
<i>Pd</i>				.06	.10	.11	-.02
<i>Pa</i>	.55		-.47	.17	.33	.13	-.01
<i>Pt</i>	-.72	.36	.39	-.25	-.25	.04	.33
<i>Sc</i>	.48			.44	.58	.44	.43
<i>Ma</i>				.01	.07	.01	-.01

Inspection of Table 3 indicates that idealized individual I represented judges who were positively correlated with the regression weights for *Sc* and *Pa* and negatively correlated with *Pt*, *Hy*, *D*, and *Hs*. For purposes of comparison, the middle section of Table 3 presents the standardized regression coefficients of the eleven scales in predicting the *actual* criterion. Goldberg (1965) found that the best predictor of the actual criterion was a simple combination of five MMPI scales: $L + Pa + Sc - Hy - Pt$. Idealized

individual I was correlated with four out of the five most valid scales in the appropriate direction of criterion prediction. In addition, it can be seen that idealized individual I was negatively correlated with *Hs* and *D*, scales of practically no validity. With the exception of these latter two scales which were overweighted, the most valid idealized individual tended to correlate with the most valid scales.

Idealized individual II, representing invalid judges, was *positively* correlated with the regression weights for *Pt*, a scale of relatively large *negative* validity. Hence a potentially valid scale was weighted in the improper direction. In addition, idealized individual II was significantly correlated with *D*, *Hs*, and *F*, three scales with practically no validity. With the exception of *Pt* which was correlated with idealized individual II in the reverse direction to its validity, idealized individual II was not related to any of the scales entering into Goldberg's best set of five scales. Idealized individual III, representing the less reliable judges, was significantly correlated with *Pt*, *Pa*, and *Hy* in the reverse direction to these scales' validities. This idealized judge also overweighted *Hs*. Although both idealized individuals II and III had a tendency to be invalid, their invalidity would appear to stem from variability in the weighting of quite different MMPI scales.

It is important to note that the correlations between idealized individuals and scale regression weights are *not* a direct test of the scales utilized or weighted in the actual judgment task. For example, if a scale regression weight does *not* correlate significantly with an idealized individual this could be due to two reasons: (a) the marker judges do not weight the scale; or (b) all of the judges (marker and nonmarker) weighted the scale, and hence the regression weight for the scale lacked sufficient variability for a given idealized individual. On the other hand, even if a scale does correlate significantly with an idealized individual, this correlation alone would not be an indication of the actual magnitude of the regression weight attached to the scale by the marker judges; such a correlation would only indicate the relative variability of judges' projections on an idealized individual with respect to the regression weights for that scale.

In light of the foregoing considerations, the regression model for the *single* judge marking each idealized individual was examined.

These were the actual judges through which the rotated factors were passed, and as such these judges would be considered an adequate representation of the corresponding idealized individual. By the examination of these single judges it is possible to determine directly the judge's weighting system for the eleven MMPI scales. The right hand side of Table 3 presents the standardized regression weights for each marker judge for the eleven MMPI scales regressed onto the 861 MMPI judgments. These weights indicate the relative emphasis a given judge placed on each scale in making his judgments of psychosis and neurosis for the profiles. The scale regression weights for the actual criterion provide a basis for determining the adequacy of each judge's model.

Marker judge I, who was among the most valid judges, correctly weighted four out of the five most valid scales (Goldberg, 1965). In addition, he tended to slightly overweight *D* and *Hs*. Marker judge II, the least valid but moderately reliable judge, correctly weighted *Sc*, and *Pa*. However, this judge tended to underweight *Pt*, *Hy*, and *L*, while overweighting *F*, *K*, *D*, and *Pd*. Although Marker judge II correctly weighted two out of the five best scale predictors, his invalidity presumably stemmed from his inappropriate weighting of seven scales. Marker judge III, an invalid and unreliable judge, correctly weighted *Sc*, although he weighted *Pt*, a negatively valid scale, with a fairly large positive weight. He tended to overweight *Hs* and underweight *L*, *Hy*, and *Pa*. The most obvious source of invalidity for this judge was his positive weighting of *Pt*, a scale of considerable negative validity. The unreliability of Marker judge III would also be another possible source of invalidity.

It is of interest to compare the models of the marker judges with the corresponding idealized individual correlations with respect to the scale regression weights. Both idealized individual I and Marker judge I present a similar picture. That is, all of the judges' projections on idealized individual I were significantly correlated with just those scales which were heavily weighted by Marker judge I. This comparability between idealized individual and Marker judge did not hold, however, for idealized individuals II and III. For example, both Marker judges II and III placed heavy emphasis on *Sc*. However, *Sc* did not correlate significantly with either idealized individual II or III. Similarly, some scales which did correlate significantly with the idealized individuals II and III did not yield

large regression weights for the corresponding marker judges. Possible noncomparability between marker judge and idealized individual can arise from the fact that idealized individual correlations can be attributed to two, confounded sources: (a) the weighting of the scale, and (b) the variability of the scale. As noted previously, the idealized individual correlations simply indicate the *relative* variability of scale weighting for all of the judges' projections on a given idealized individual; the models for the marker judges yield the actual magnitude of the scale weights for a single judge marking an idealized individual.

Discussion

In comparing the present data with those of Horn and Stewart (1968), recall that two of Horn's three subject factors were totally uninterpretable, and Horn's varimax subject factors did not pass through real individuals in the factor space of subjects. The present data point out the usefulness of identifying idealized individuals on the basis of judgmental and personological variables, as well as realistically representing their judgment strategies by considering the models for the marker judges. It could even be argued that rather than treating each of the 29 judges separately in examining their judgmental strategies, only three idealized types need be studied. Thus, given a particular constellation of judgmental and personological correlates, it would be possible to predict an idealized judge's model for a set of new data, provided that the idealized judge exhibited a similar pattern of correlations with the judgmental and personological measures.

This is not to say, however, that idealized individuals would necessarily emerge in a new study with judgment models identical to those of the present study. This should only occur when, in fact, the pattern of outside correlates is identical to those found in the present study, an unlikely finding in a new study. However, it can reasonably be predicted that new data should exhibit individual differences in judgmental viewpoints for judgments of psychosis vs. neurosis from the MMPI. Further, it would be predicted that three major judgmental variables would distinguish the different perceptual viewpoints: validity, predictability and reliability. Should different types of invalid judges emerge, the weights they place on the MMPI scales as well as their relative reliability

should also distinguish among them. If a rotation of the subject factors could be found which replicated the present pattern of judgmental and personological correlates, the judgment models for the idealized types should also replicate. These hypotheses are currently being tested on data similar to those used in the present study. Since the present data consisted of a fairly homogeneous group of judges (Minnesota-trained clinicians), a more heterogeneous sample might well lead to more than three idealized judges.

Although most of the variables used in the present study were primarily judgmental variables based on the same data which were factor analyzed, the present technique suggests a means whereby types of clinical judges might be identified a priori to performing the judgment task. It is suggested that future research be directed to uncovering a variety of personological and intellectual correlates of the idealized judges. If constellations of characteristics associated with different types of clinical judges were isolated, the possibility exists for a priori typological identification for any given judge.

REFERENCES

- Cliff, N. The 'idealized individual' interpretation of individual differences in multidimensional scaling. *Psychometrika*, 1968, 33, 225-232.
- Goldberg, L. R. Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs*, 1965, 79, (9, Whole No. 602).
- Goldberg, L. R. Simple models or simple processes? Some research on clinical judgments. *American Psychologist*, 1968, 23, 483-496.
- Goldberg, L. R. The search for configural relationships in personality assessment: The diagnosis of psychosis vs. neurosis from the MMPI. *Multivariate Behavioral Research*, 1969, 4, 523-536.
- Goldberg, L. R. Man vs. model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin*, 1970, 73, 422-432.
- Hammond, K. R., Hursch, C. J., and Todd, F. J. Analyzing the components of clinical inference. *Psychological Review*, 1964, 71, 438-456.
- Hoffman, P. J. The paramorphic representation of clinical judgment. *Psychological Bulletin*, 1960, 57, 116-131.
- Horn, J. L. and Stewart, P. On the accuracy of clinical judgments. *British Journal of Social and Clinical Psychology*, 1968, 7, 129-134.
- Meehl, P. E. A comparison of clinicians with five statistical methods

- of identifying psychotic MMPI profiles. *Journal of Counseling Psychology*, 1959, 6, 102-109.
- Messick, S. and Kogan, N. Personality consistencies in judgment: Dimensions of role construct. *Multivariate Behavioral Research*, 1966, 1, 165-175.
- Pederson, D. M. The measurement of individual differences in perceived personality trait relationships and their relation to certain determinants. Unpublished doctoral dissertation, Urbana: University of Illinois, 1962.
- Skager, R. W., Schultz, C. B., and Klein, S. P. The multidimensional scaling of a set of artistic drawings: Perceived structure and scale correlates. *Multivariate Behavioral Research*, 1966, 1, 425-436.
- Snyder, F. R. and Wiggins, N. Affective meaning systems: A multivariate approach. *Multivariate Behavioral Research*, 1970, 5, 453-468.
- Tucker, L. R. Intra-individual and inter-individual multidimensionality. In H. Gulliksen and S. Messick (Eds.), *Psychological Scaling: Theory and Applications*. New York: Wiley, 1960, pp. 155-167.
- Tucker, L. R. A suggested alternative formulation in the developments by Hirsch, Hammond and Hirsch, and by Hammond, Hirsch, and Todd. *Psychological Review*, 1964, 71, 528-530.
- Tucker, L. R. Learning theory and multivariate experiment: Illustration by determination of generalized learning curves. In R. B. Cattell (Ed.), *Handbook of Multivariate Experimental Psychology*, Chicago: Rand McNally, 1966, 476-501.
- Tucker, L. R. and Messick, S. An individual differences model for factor-analytic techniques." *Psychological Bulletin*, 1968, 70, 345-354.
- Tucker, L. R. and Messick, S. An individual differences model for multidimensional scaling. *Psychometrika*, 1963, 28, 333-367.
- Walters, H. A. and Jackson, D. N. Group and individual regularities in trait inference: A multidimensional scaling analysis. *Multivariate Behavioral Research*, 1966, 1, 145-163.
- Wiggins, N. Individual viewpoints of social desirability. *Psychological Bulletin*, 1966, 66, 68-73.
- Wiggins, N. and Fishbein, M. Dimensions of semantic space: A problem of individual differences. In J. R. Snider and C. E. Osgood (Eds.), *The Semantic Differential Technique*. Chicago: Aldine Publishing Co., 1969, pp. 183-193.
- Wiggins, N. and Hoffman, P. J. Three models of clinical judgment. *Journal of Abnormal Psychology*, 1968, 73, 70-77.
- Wiggins, N., Hoffman, P. J., and Taber, T. Types of judges and cue-utilization in judgments of intelligence. *Journal of Personality and Social Psychology*, 1969, 12, 52-59.
- Wiggins, N. and Wiggins, J. S. A typological analysis of male preferences for female body types. *Multivariate Behavioral Research*, 1969, 4, 89-102.

A SPECIAL REVIEW OF BUROS' PERSONALITY TESTS AND REVIEWS

FRED DAMARIN

The University of Delaware

OSCAR BUROS has reprinted the reviews of all the personality instruments that appear in his *Mental Measurements Yearbooks* together with significant additional material in a single new volume called *Personality Tests and Reviews*.¹ The result is a Domesday Book for those of us who are interested in personality studies. The original Domesday Book was a survey of England's estates and manors that William the Conqueror ordered in order to consolidate his administrative authority. Buros' *Personality Tests and Reviews* (PTR) is a survey of the copyrighted personality instruments that are the major real properties in the realm of personality studies. It records often conflicting estimates of their worth and is a source of information for those who wish to utilize them. The original Domesday was incomplete; it ignored London. PTR is also incomplete; it ignores uncopyrighted devices like the California *F* scale.² Despite this incompleteness it is, like its prototype, an effective instrument of social control.

¹ Excerpted MMY reviews account for about 1100 of the 1700 pages in this book. There are 555 pages of supplementary indices including a 260 page Test Index that describes in-print and out-of-print nonprojective instruments and in-print and out-of-print projective instruments in separately alphabetized sections. It also serves to up-date lists of references from the sixth MMY. Other indices in this book list personality instrument titles, authors, publishers, and reviewers. There is a Classified Index for all tests (of whatever sort) in any MMY and for all the book reviews excerpted there. Buros reprinted the APA-AERA-NCME *Standards for Educational and Psychological Tests and Manuals* and last, but not least, he contributed an illuminating 17 page preface to the whole works that includes many valuable statistical tables.

² Our sample of copyrighted instruments is assumed to be a numerically large fraction of the total population of personality devices and is assumed to be unbiased in all important respects. No convenient way of checking this assumption has occurred to us.

Social control has never been all that popular: The Saxon gentry dubbed William's book Domesday (Doomsday) because it posed a record from which there was no appeal. In some quarters Buros' efforts have occasioned something like the same dismay. In the *Second Mental Measurement's Yearbook* (MMY) Buros reprinted a selection of letters that bitterly criticized the first MMY and he reported that some test publishers were reluctant to forward samples for review. By now the MMY series is so prestigious that Buros has little difficulty in securing the cooperation of publishers. But active resentment may merely have mellowed into benign neglect for Buros' whole undertaking seems never to have received foundation support and his own labors may never have been properly acknowledged by the educational and psychological communities.

One way to acknowledge Buros' contribution is to buy his book. Since it is expensive, there will be a temptation to ask one's local librarian to buy the book instead, in case anyone wants to read up on a particular instrument. PTR is unquestionably useful for this purpose. Teachers and research workers in the area of personality studies may find, however, that it has other uses that will amply repay the cost of ownership. Buros' pages present an implicit history of personality assessment that is richer and more suggestive than any other the writer has come upon. This review will attempt to make parts of this implicit history explicit in order to assess the future prospects of such assessment. For this purpose it is necessary to structure the data in Buros' book. The reader is warned that a closer acquaintance with the work might suggest different structuring principles to him and that these could lead to different and possibly less pessimistic conclusions.

Exponential Growth

In surveying this book it is helpful to ask questions. The writer wanted to know whether (and in what sense) there has been progress in the art of personality measurement in the last 25 or 30 years. As a first step it seemed desirable to look at the rates at which copyrighted measuring devices and new publications on these have accumulated over the years. Since the PTR Test Index gives the date at which each instrument was copyrighted it was possible to make cumulative frequency distributions over years. This was done

separately for projective and nonprojective instruments. Table 1 displays the total number of each sort of device copyrighted up to 1925, up to 1930, and up to the end of each subsequent five year interval through 1965.³ Figures for 1966-1968 are given, but these may be incomplete and the figures for 1970 are projections from trends established up through 1965. These data appear in column 2 and 3 of Table 1. Columns 4 and 5 record the cumulative number of references available for the nonprojective and projective measures separately beginning in 1940. These figures were calculated from a frequency table that appears on page xxiii of Buros' introduction to PTR. The last two columns in our Table 1 display ratios of references to instruments (separately in the nonprojective and projective areas) for each data point since 1940.

The growth trends for instruments have been graphed in Figure 1, the trends for references appear in Figure 2, and the trends for references-per-instrument appears in Figure 3. All of these graphs are semi-logarithmic so that the scale of years on the X axis is linear while the scale of cumulative instruments, cumulative refer-

TABLE 1

Cumulative Instruments Copyrighted, Cumulative References, and Ratios of References to Instruments

Year (1)	Tests-Copyrighted		References		References/Test	
	Nonproj. (2)	Proj. (3)	Nonproj. (4)	Proj. (5)	Nonproj. (6)	Proj. (7)
(1970)	(510)	(95)	(13,000)	(8,500)	(25.5)	(93.6)
1968	405	93	10,947	7,753	27.0	83.4
1965	369	89	8,365	6,786	22.7	76.2
1960	301	78	4,942	5,191	16.4	66.6
1955	223	58	3,119	3,609	14.0	62.2
1950	179	37	2,031	1,724	11.3	46.6
1945	139	16	1,214	653	8.7	40.8
1940	100	10	714	225	7.1	22.5
1935	44	4	—	—	—	—
1930	9	4	—	—	—	—
1925	2	2	—	—	—	—

Note.—Figures for 1966-1968 may be incomplete. Data given for nonprojective (nonproj.) and projective (proj.) instruments at intervals since 1925. Parenthesized values are crude extrapolations from recent growth rates. Dashes indicate missing dates. All information taken from Buros (1970)

³ The data in this review on the total number of instruments available at any given time need not always agree exactly with comparable figures in PTR. Buros seems to count scoring services as separate instruments but this review does not.

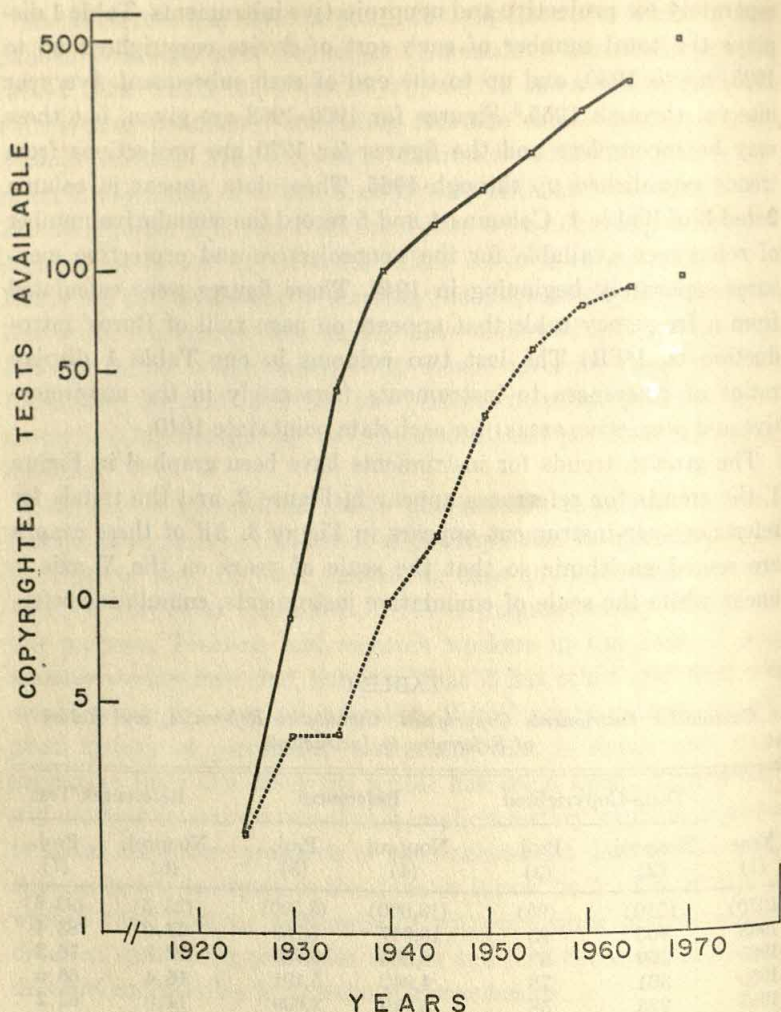


Figure 1. Growth curves for personality instrument copyrights. The logarithm of the total number of copyrights issued is plotted against the date at five year intervals from 1925 to 1965. The solid line represents non-projective devices while the dashed line represents projective measures. The points representing 1970 are crude extrapolations. The data are from Buros (1970).

ences, and ratios on the three Y axes are logarithmic. Markedly linear trends—with some inflection points—are visible in all three graphs. Over long periods of time the number of instruments, the number of references, and the number of reference *per* instrument

tend to grow exponentially: They double every few years. Over shorter periods of time there are often perturbations that are not displayed in figures 1, 2 or 3. During and after the second World War, for example, the production of both instruments and re-

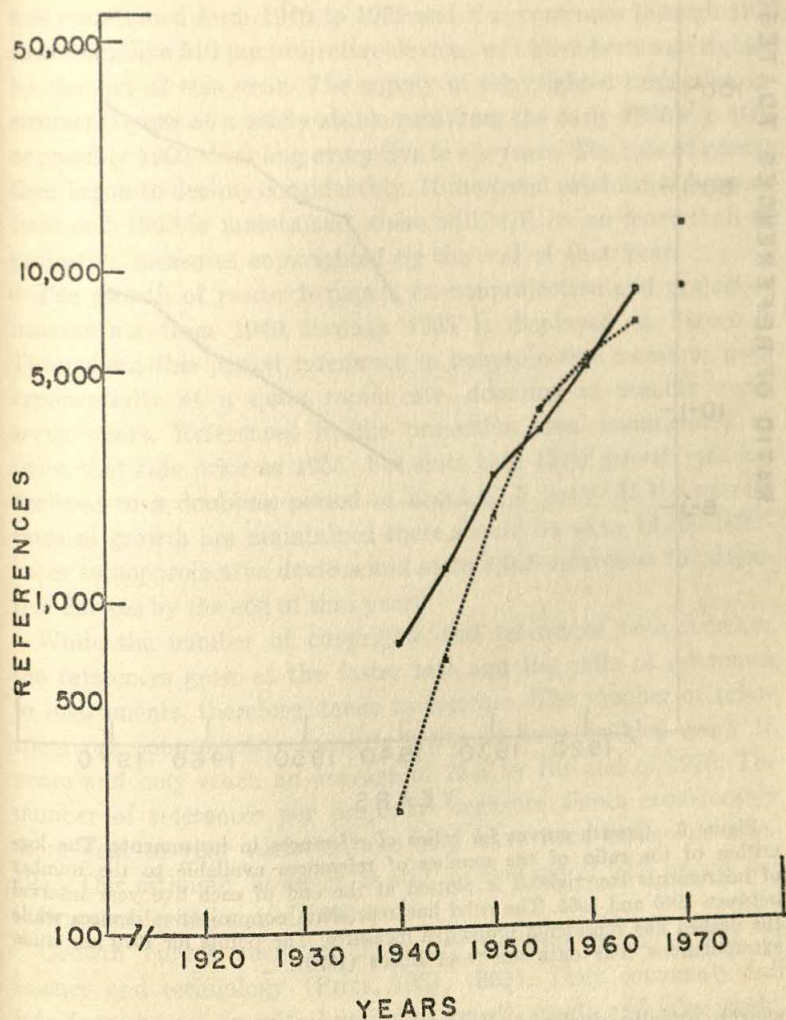


Figure 2. Growth curves for personality instrument references. The logarithm of the total number of references available is plotted against the date at five year intervals from 1940 to 1965. The solid line represents references to nonprojective devices while the dashed line represents references to projective measures. The points representing 1970 are crude extrapolations. The data are from Buros (1970).

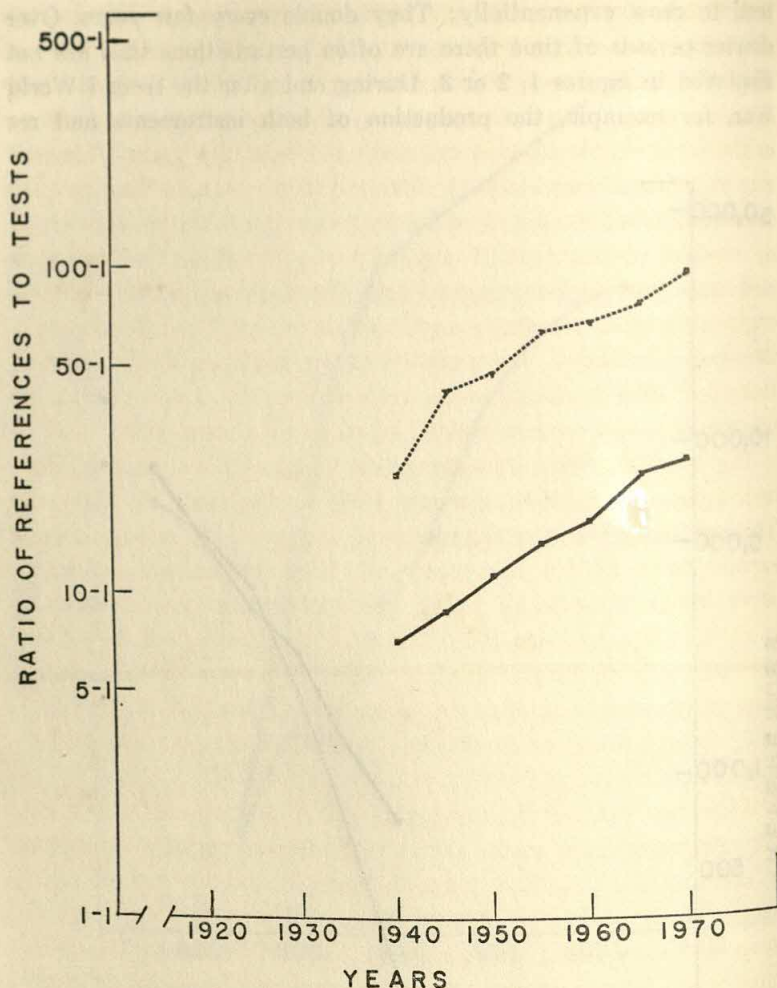


Figure 3. Growth curves for ratios of references to instruments. The logarithm of the ratio of the number of references available to the number of instruments copyrighted is plotted at the end of each five year interval between 1940 and 1965. The solid line represents nonprojective devices while the dashed line represents projective measures. The points for 1970 are crude extrapolations. The data are from Buros (1970).

search articles declined and then rose very sharply as though a deficit were being made up. Growth then resumed as if nothing had happened.

Concentrating on the long term trends in these data, we note that the most rapid growth in new copyrights occurred in the nonpro-

jective area between 1925 and 1935. The total supply of these instruments then doubled about every 2.25 years! Between 1935 and 1940 this rate of copyrighting declined and after 1940 the rate declined again to a relatively stable value that produces a doubling in the total supply of such instruments every 12.5 years. This rate was maintained from 1940 to 1965 and if it continues through 1970 something like 510 nonprojective devices will have been copyrighted by the end of this year. The supply of copyrighted projective instruments grew at a fairly stable rate from the early 1930's to 1955 or possibly 1960, doubling every five to six years. The rate of growth then began to decline considerably. If the trend established between 1960 and 1965 is maintained, there will still be no more than 95 projective measures copyrighted by the end of that year.

The growth of research papers on nonprojective and projective instruments from 1940 through 1965 is displayed in Figure 2. Throughout this period references to nonprojective measures grew exponentially at a quite rapid rate, doubling in number every seven years. References in the projective area accumulated at twice that rate prior to 1955; but since then their growth rate has declined to a doubling period of about 11.5 years. If the current rates of growth are maintained there should be some 13,000 references to nonprojective devices and some 8,500 references to projective devices by the end of that year.

While the number of copyrights and references both increase, the references grow at the faster rate and the ratio of references to instruments, therefore, tends to increase. The number of references per nonprojective measure seems to have doubled every 15 years and may reach an average of 25.5 by the end of 1970. The number of references per projective measure shows considerably more year to year variation and may have been in decline since 1955. Our estimate, however, is still an average of 93.6 papers per instrument by the end of 1970.

Growth curves such as these are familiar to historians of science and technology (Price, 1961, 1962). They commonly fall into four phases: an initial period of growth spurts and false starts, a period of steady exponential growth, a decline into merely linear growth, and finally either stagnation or renewal if new techniques or points of view revive the field and bring about a new period of exponential growth. Nonprojective personality instruments are

still at the second of these four stages, for all the relevant growth rates are still exponential. The projective measures seem well embarked on the third leg of this four stage journey, for the rate of copyrighting and researching them has declined so steadily that there are now less than half the number of new instruments and publications that the earlier growth rates (prior to 1955) would have led one to expect. Unless new techniques or insights appear that can revive this field, the whole projective movement may be moribund by 1980. The seriousness of this prospect becomes apparent when we realize that the label "projective test" covers a great variety of devices while the nonprojective instruments consist chiefly of questionnaires.⁴ What will happen to the field of personality studies if, after a half century of work, the self-report questionnaire emerges as the only successful personality measuring device?

Quantity and Quality

Since concern about these quantitative issues would abate if the quality of personality measurement were clearly improving, it is worth examining a number of mechanisms by which increments in quality might be produced. Progress is sometimes said to stem from competition and the survival of the fittest, from research, and from the application of higher professional standards. The data in PTR may be used to evaluate each of these hypotheses.

Competition and Progress

The continuous arrival of new personality measures and research studies leads us to suppose that the older personality measures are just as continually falling by the wayside, obsolescing considerable portions of the early literature as they go. This view gains credence from studies of the changing popularity of personality instruments in clinics and business organizations. Sundberg (1961), for example, collected usage data in several sorts of clinical services in 1959 for comparison with similar data from Louttit and Browne

⁴ Nonprojective devices need not consist chiefly of questionnaires. As Cattell and Warburton (1967) demonstrate, there are almost as many types of tests in Cattell's Objective Analytic Test Battery as there are in all of PTR. But this nonprojective battery was copyrighted in 1955, it has accumulated only 23 references, and it is now out of print. We understand that this state of affairs is temporary and hope to see further work done on these measures for the data of this review highlight their strategic importance.

(1947) covering 1946 and 1935. Sundberg found that between 1935 and 1946 there was a turnover rate of 60 per cent in the twenty most favored measures. Between 1946 and 1959 this rate was 38 per cent. While these changes are dramatic, the figures confound changes in the perceived suitability of instruments for particular diagnostic goals—such as assessing brain damage—with changes in the diagnostic goals themselves. As Sundberg so clearly points out, changes of this latter sort prevailed throughout clinical psychology between 1935 and 1959. The rapid turnover of instrument preferences during this era is, therefore, no guarantee that direct competition between devices occurred or that progress in instrument design was being made.

The tendency for instruments to go out of print might be a more nearly optimal index of competition and hence of technological progress in testing. We would expect, for example, that most of the recently copyrighted measures would still be in print but that those that were copyrighted at progressively earlier dates would have suffered progressively more attrition, so that by now nearly all the earlier devices would have disappeared from the market. The Text Index in PTR lists in- and out-of-print instruments in separate sections and it was easy to use this information to construct Table 2. Column 1 of this table subdivides over 40 years of measurement

TABLE 2

Personality Instruments Copyrighted and Instruments Out-of-Print in 1970

Interval of Years (1)	Test Copyrighting Frequency		Out of Print in 1970		Proportion Now Out of Print	
	Nonproj. (2)	Proj. (3)	Nonproj. (4)	Proj. (5)	Nonproj. (6)	Proj. (7)
1966-68	35	4	1	0	.03	.00
1961-65	68	11	1	0	.02	.00
1956-60	78	20	10	1	.13	.05
1951-55	44	21	8	4	.18	.19
1946-50	40	21	11	4	.28	.19
1941-45	39	6	17	2	.44	.33
1936-40	56	6	43	2	.77	.33
1931-35	35	0	25	0	.71	.00
1926-30	7	2	4	0	.57	.00
Up to 1925	2	2	1	0	.50	.00
Totals	404	93	121	13	.30	.14

Note.—Figures for 1966-1968 may be incomplete. By consecutive half decade intervals. Data abstracted from Buros (1970).

history into blocks of (usually) five years each. Columns 2 and 3 record the incidence of nonprojective and projective instrument copyrights *within each time block*. Columns 4 and 5 report the number of these devices that are listed as out of print as of the publication of PTR. The final two columns give the ratios of out-of-print to total instruments within each block of time.

These out-of-print ratios do behave in certain respects as though they measured obsolescence: They increase as one goes backward in time, at least until the middle 1930s. The index for projective measures is uniformly lower than for nonprojective measures, suggesting that the latter are more likely to go out of print. Since the rate of copyrighting nonprojective measures is still exponential, the continual arrival of new devices may indeed put the older ones out of business.

As one looks back beyond the middle 1930s, however, the out-of-print ratios in the last two columns stop rising and begin to decline. Some very early instruments seem to have a competitive advantage over later ones, but this is not what conventional notions of obsolescence would lead one to expect. Further difficulties arise when one examines the references attached to the in- and out-of-print devices; the out-of-print versions seem to have less than their fair share. The 30 per cent of all nonprojective measures that have gone out of print account for only 7 per cent of the literature on nonprojective instruments (774 references out of 10,947). The 14 per cent of all projective measures that are now out of print account for a little over 1 per cent of their literature (90 out of 7,753 references). Looking at these data in another way, we find that while a few inventories such as Thurstone's Personality Schedule or the Bogardus Social Distance Scales went out of print with over 50 references, the majority departed with far less. The nonprojective measures that are still in print average 36 references apiece whereas those that are out of print average only 6.4 references and have a median of two! The projective devices in print average 95.8 references apiece but out-of-print projective measures average 6.9 references and have a median of 5.

If competition implied progress, personality measures ought to go out of print because they are dated by technically more advanced *successors*. An undetermined but probably large number of personality measures may go out of print because they cannot compete with

well entrenched *predecessors*. This may occur because good ideas are widely copied and because the competition is between instruments that are both original and sound and a host of inferior imitations that are gradually forced off the market. On the other hand, the demand for innovation in personality measurement may be so inelastic that modest improvements have little chance of acceptance. In neither case does competition imply automatic progress and there is, therefore, room to suspect that the field of personality measurement may actually be in a state of technological stagnation.

Research and Progress

The sheer volume of research that is being done with personality instruments might seem to guarantee progress in design, at least until we glance at some data in a table that Buros provides on page xviii of his introduction to PTR. This table shows that the first six MMY's recorded a total of 2,802 tests of all sorts which fell into 15 major categories. The category of personality measures contained 386 members or about 14 per cent of the total. The entire collection had a bibliography of 23,763 titles to which the category of personality measurement contributed 11,214, or almost 50 per cent of the total. The next largest contributors were two measurement categories that are often absorbed into expanded definitions of personality. There were 290 intelligence tests with 5,494 references and 365 vocational measures with 2,650 references. In marked contrast the tests in such academic areas as Business Education, English, Foreign Languages, Mathematics, Science, and Social Studies constituted a pool of 1058 instruments with 1194 references, which is only a little more than one reference per test. Almost everyone would agree that academic tests are more valid as a group than the personality measures and yet research publications favor the latter by an enormous margin. There is, evidently, a great deal of difference between research *on* tests and research *with* tests and most of the references in the personality area probably fall into the latter category.

Personality researchers may view their instruments as tools that are important for the *kind* of service they render rather than for the *quality* of that service. Zipf (1949) suggests that the most used members of any set of tools will tend to be versatile. There are at least two senses in which tools can be versatile, however.

They can perform many different functions, like a jackknife, or they can perform one service that is involved in many different kinds of work, like a hammer. Since the number of scores provided by an instrument might index its jackknife-type versatility, we asked whether the more multivariate instruments tended to accumulate references more rapidly.

In testing this hypothesis we turned to a table provided by Buros on page xxiv of his introduction to PTR. This table gives the number of references for each of those 89 personality measures that have more than 25 references apiece. Some of the top measures on this list could have accumulated many references merely because they have been in circulation since the 1930s. In order to correct for this the out-of-print devices were dropped and the references for each in-print instrument were divided by the number of years that elapsed between the year of its first copyright and 1970. These in-print instruments were then listed in descending order of their publication rates in Table 3. Information on the number of scores provided by each measure was then sought in the PTR Test Index. The data on projective measures proved incomplete; in many cases the number of scores was not even listed, perhaps because they are thought capable of measuring an indefinitely large number of traits. In any event it was possible to test the versatility hypothesis only on non-projective instruments.

The 62 in-print, nonprojective measures were dichotomized at the median on the "references" variable to give 31 devices with more than 3.20 references per year and 31 with less. The "number of scores" variable was dichotomized as near its median as possible to give 35 devices with six or more scores and 27 with five scores or fewer. Twenty-two of the more frequently published measures had six or more scores. Chi Square for this fourfold table is 5.30 and with a two-tailed test the associated probability is about .025. The corresponding contingency coefficient is .29, the phi-coefficient is .25, and the tetrachoric correlation was estimated as .45. There appears to be a modest relationship between multivariate versatility and popularity as a research instrument—at least among that group of instruments that have accumulated 25 publications or more.

There are measures in Table 3 that are being vigorously researched even though they provide relatively few scores. Several

of these appear to assess the "first factor" of the MMPI, the one that Edwards (1970) calls social desirability. This dimension probably displays our second sort of versatility, for it is involved in many research problems especially when questionnaires are used to identify psychopathology. Almost all the symptoms of various sorts of psychopathology are socially undesirable and will be confessed only by those who are capable of making negative statements about themselves. Questionnaire measures of psychopathology, therefore, tend to produce a common factor even though the syndromes being measured may have very little in common.

There is, of course, no necessary relationship between versatility and validity. One response dimension should not be used to index six or seven conceptually different syndromes and multivariate instruments that utilize a single measurement method (typically the questionnaire) probably provide less validity per trait than would an equally comprehensive battery of univariate scales in which each trait is measured by a method that is maximally appropriate to it.

Standards and Progress

Personality testers, particularly projective testers, are often urged to study professional test standards and apply them to their own wares. Buros reprints the current APA-AERA-NCME *Standards for Educational and Psychological Tests and Manuals* in PTR, but he also calls our attention to certain insufficiently appreciated peculiarities in this document. It focuses on reporting the manual, not on instrument design. The *Standards* (and the earlier *Technical Recommendations*, 1954) assert that projective instruments pose special problems because they have both nomothetic and ideographic aspects and they urge that their authors report the nomothetic aspects fully in their manuals. Many authors may have accepted this advice as an invitation to force data from their instruments into something like a manual. The results, to judge from Buros' reviewers, is often a laboriously intricate scoring system with statistically undesirable properties including low reliability. Other authors may try to capitalize on the supposedly ideographic aspects of their instruments by trying to show that experts can use the material to make interesting predictions. The

TABLE 3

Personality Instruments with 25 or More References

Tests (and rankings)	References per year	Scores
Minnesota Multiphasic Personality Inventory (1)	88.4	14 plus
Rorschach (1*)	76.5	?**
Edwards Personal Preference Schedule (2)	41.0	15
Thematic Apperception Test (2*)	36.4	?**
California Psychological Inventory (3)	28.4	18
Maudsley Personality Inventory (4)	24.5	2
Sixteen Personality Factor Questionnaire (5)	18.0	20
The Guilford-Zimmerman Temperament Survey (6)	14.5	10
Bender-Gestalt Test (3*)	13.4	?**
Study of Values (7)	12.2	6
Rosenzweig Picture-Frustration Study (4*)	11.9	15
Machover Draw-A-Person Test (5*)	10.5	?**
Personality and Personal Illness Questionnaires (8)	9.7	16
The Holtzman Ink Blot Technique (6*)	9.3	22
The Personality Inventory (Bernreuter) (9)	8.7	6
California Test of Personality (10)	7.8	15
Eysenck Personality Inventory (11)	7.6	3
Interpersonal Check List (12)	7.3	8 plus**
Stern Environment Indexes (13)	6.5	48
Omnibus Personality Inventory (14)	6.4	15
Survey of Interpersonal Values (15)	6.0	6
The Blacky Pictures (7*)	5.9	13**
The Adjective Check List (16.5)	5.7	24
H-T-P (8*)	5.7	?**
The IPAT Anxiety Scale Questionnaire (16.5)	5.7	6
Multiple Affect Adjective Check List (18)	5.6	3
Szondi Test (9*)	5.6	8
Inpatient Multidimensional Psychiatric Scale (19)	5.1	10
College and University Environment Scales (20.5)	5.0	7
Spiral Aftereffect Test (20.5)	5.0	1**
The Adjustment Inventory (Bell) (22)	4.8	6
Mooney Problem Check List (23)	4.7	11
Structured-Objective Rorschach Test (10*)	4.1	15
Embedded Figures Test (24)	4.0	1**
Vineland Social Maturity Scale (25)	3.9	1**
Rotter Incomplete Sentences Blank (11*)	3.8	1**
Gordon Personal Profile (26)	3.8	4
Cornell Medical Index-Health Questionnaire (27)	3.6	1**
The FIRO Scales (28)	3.5	6 plus
Stern Activities Index (29)	3.4	48
Personal Orientation Inventory (30)	3.3	12
Shipley-Institute of Living Scale for Measuring Intellectual Impairment (31)	3.2	4
Stanford Hypnotic Susceptibility Scale (32)	3.2	1**
Goldstein-Scheerer Tests of Abstract and Concrete Thinking (33)	3.1	5**
The Hoffer-Osmund Diagnostic Test (34)	3.1	5
The Guilford-Martin Inventory of Factors GAMIN (35)	3.1	5
Kahn Test of Symbol Arrangement (12*)	3.0	?**
Jr.-Sr. High School Personality Questionnaire (36)	3.0	14

TABLE 3 (Continued)

Tests (and rankings)	References per year	Scores
The IES Test (13*)	2.9	14
Minnesota Counseling Inventory (37)	2.8	9
Cornell Index (38)	2.7	1**
Vocational Preferences Inventory (39)	2.6	11
Children's Apperception Test (14*)	2.6	7**
The Guilford-Martin Personnel Inventory (40)	2.6	3
Lowenfeld Mosaic Test (15*)	2.5	7**
Myers-Briggs Type Indicator (41)	2.4	4
STS Youth Inventory (42)	2.4	8 plus
Thurstone Temperament Schedule (43)	2.3	7
An Inventory of Factors STDCR (44)	2.3	5
Make A Picture Story (16*)	2.2	7**
Tennessee Self Concept Scale (45)	2.2	30
Welsh Figures Preference Test (46)	2.1	27
The Humm-Wadsworth Temperament Scale (47)	1.9	47
Kent-Rosanoff Free Association Test (17*)	1.9	7**
Activity Vector Analysis (48)	1.9	6
Interpersonal Diagnosis of Personality (18*)	1.9	32**
Gordon Personal Inventory (49)	1.9	4
Memory-For-Designs Test (50)	1.8	1**
It Scale for Children (51)	1.8	1**
Concept Formation Test (Vigotsky) (52)	1.6	1**
A-S Reaction Study (53)	1.6	1**
Babcock Test of Mental Efficiency (54)	1.5	10
KD Proneness Scale and Check List (55)	1.4	2**
Social Intelligence Test (56)	1.3	6
Kuder Preference Record-Personal (57)	1.3	6
The Empathy Test (58.5)	1.3	1**
The Purdue Master Attitude Scales (58.5)	1.3	9 plus
Attitude Interest Analysis Test (60)	1.2	1**
Security-Insecurity Inventory (61)	1.0	1**
Personal Adjustment Inventory (62)	0.7	5

Note.—Numbers in parentheses following title give rank of instruments in terms of references accumulated per year since first being copyrighted. Ranks were assigned separately for nonprojective and projective measures. Column two contains average number of references per year. Column three contains numbers of scores per instrument. Question marks denote devices where this is especially ambiguous. All data taken from Buros (1970).

* projective instrument.

** Number of scores not given in Personality Test Index in Buros (1970).

results are usually one more demonstration of the fallability of experts, especially when they are competing with a regression equation.

Hindsight now suggests that the real problem is that some instruments are composed of items whereas others are not (DuBois, 1970). When independent, dichotomously scored units can be grouped in various ways to form scales it is relatively easy to upgrade the test manual. One can provide more adequate norms, report more extensively on reliability, disclose new correlations

between scales and external criteria. A stricter application of the current test standards should produce more of these benefits and even lead indirectly to some improvements in the instrument itself. The author might decide to add items to some scales to bring their reliabilities up to the competition.

Many projective measures do not possess items in the conventional sense. Complex patterns of behavior may be evoked, but the components are often experimentally or statistically interdependent and they may be difficult to group into scales in ways that satisfy psychometric criteria. Improving an instrument of this type may require changes in its design. The solution can be as simple as Holtzman's conversion of the classical Rorschach into a format with one response per ink blot. Alternatively, one may have to experiment with the device, modifying the instructions and the stimuli, until the elicited behavior is well enough understood to recast the entire procedure into a format with greater psychometric utility.

Since this sort of instrument-oriented research may receive very little encouragement in any quarter, there is a good chance that very little of it is being done. There is no reason to suppose that stricter application of the manual-oriented *Technical Recommendations* and *Standards* will improve matters. It may instead accelerate the abandonment of a wide variety of formats—notably but not exclusively the projective techniques—in the mistaken idea that they are unsound when they are merely undeveloped. It may lead to the acceptance of questionnaires as psychologically sound when they are advanced only from the psychometric viewpoint.

The Costs of Stagnation

PTR presents personality measurement in a new and unexpected light: There is considerable activity in some quarters but in others there is evidence of impending stagnation. Inferior instruments are supposedly hazardous to test takers and some serious professional thought, along with at least one Congressional investigation, has been devoted to protecting their interests. Teachers and researchers in the area of personality studies have been less concerned about the possible costs of stagnation to themselves. The researchers probably suppose that personality measures are tools that may be used indefinitely without much concern about their

maintenance and improvement. By using them in substantive research they may even suppose that they are making improvements. The exact opposite may actually be true.

Most psychologists use personality instruments to define constructs. In order to learn more about anxiety, extroversion, or need achievement they administer measures of these traits and apply statistical procedures to the scores. The results are sometimes insignificant and the experimenter must then conclude that he misunderstood his construct or that he chose a poor measure of it (or both). In a field where type II errors are as frequent as they are said to be in personality research (Cohen, 1962) these doubts will accumulate over measures and over time. Since it is easier to change one's ideas than one's tools, subsequent researchers may try to design more "insightful" studies with the old instruments instead of producing better instruments. When thousands of problematic studies accumulate, as they have on the most popular measures, a few skeptics may try to sift the data to see what is wrong, but the information may then be lost in the general glut. At some point along the way there may be questions from allied social science disciplines as to whether people who talk about personality research are really capable of advancing our knowledge of human behavior.

While technological stagnation does pose a threat to their scientific credibility, teachers and researchers in this area are by no means doomed. Their situation is serious but they possess means for improving it. They can extend and refine their analysis of personality constructs in two directions. Given any personality trait or state they could ask why it might be more accurately assessed with one sort of measuring device rather than with another. They might then ask how socially important criteria for their instruments can be provided out of the data of everyday life.

The Utility of Oscar Buros

An interest in problems of this sort should change one's view of PTR from a doomsday book to a valuable aid. Buros (1970, page xxvii) asks his reviewers to go beyond the manual, to compare tests, to say which is best, to praise good work and to censure bad. Beyond this his instructions are open ended—even projective. The reviewers decide for themselves what excellence

is and they attack his issue from many angles. PTR is suggestively rich, but its full value may be realized only if one has some means of isolating information without being distracted by irrelevant material.

Classification schemes can be used to underline design characteristics that are common to diverse personality measures. Many such classifications have been suggested (Cattell, 1957, Cattell and Warburton, 1967, Guilford, 1967, Symonds, 1931). For illustrative purposes we will sort tests into six families depending on the nature of the stimulus presented to the subject and the nature of the response expected from him. This classification may be applied to most of the instruments in PTR and it will also accommodate many of the experimental measures in Cattell and Warburton's (1967) compendium, *Objective Personality and Motivation Tests*. The scheme is illustrated in Table 4.

An Index to Controversy

"Process tracing" experiments postulate a series of stages or processes that mediate between stimulus and response (Woodworth and Schlosberg, 1958). In some of the test families in Table 5 these processes seem open to inspection, for when a subject draws a picture (Family 1) or constructs a toy world (Family 2) he produces a series of qualitatively distinctive responses that terminate in a finished product. The result may count as a single item, but all of the subject's behaviors can be referred to the consecutive stages by which this item-response came into being. While the devices in Families 3 through 6 usually contain several items, the behaviors appropriate to any one of these tend to be less overt. In the case of questionnaires we have, instead of an audible train of associations, a silent, routinized internal switching that leads to "yes" or "no." Hidden processes are likely to be poorly understood and a potential source of controversy among competing personality testers. PTR can be read as an index to these controversies and as a means of identifying some potentially important substantive research problems that are buried in the measurement literature.

The perceptual response to instruments like the Rorschach (Family 3) occurs so rapidly that there is debate about the relative importance of perceptual styles (such as color dominance) as

TABLE 4

Classification of Personality Measures

Stimulus Material	Type of Response	Examples
<i>Projective Measures</i>		
1. Blank paper	Drawings	Figure drawing devices of all sorts: H-T-P, Eight Card Redrawing Tests; Drawing Changes Under Disapproval (T-102).
2. Pieces or parts	Construction of a whole	Lowenfield Kaleidoblocks, Lowenfield Mosaic, Make A Picture Story, Picture World Test, Toy World Test; Neighbor Preferences (T-343)
3. Ambiguous visual patterns	"Imaginative" perception	Holtzman Ink Blot Technique, Howard Ink Blot Test, Rorschach, Structured Objective Rorschach; Unstructured Drawings (T-20), Unstructured Drawings Check List (T-327), Autistic Projection (T-369).
4. Pictures	Verbal associations, Semantic units or systems	Blacky Pictures, Children's Apperception Test, Thematic Apperception Test, Rosenzweig Picture-Frustration Study; Brunswik's Faces (T-92), Picture Exploration (T-104).
5. Semantic units	Verbal associations	Association Adjustment Inventory, Kent-Rosanoff Free Association Test, Sentence Completion Tests of all sorts; Association of Emotional Words (T-14), Emotionality of Comment (T-36), Sentence Completion (T-52), and others
<i>Nonprojective Tests</i>		
6. Semantic units or systems	Semantic units or systems	Psychiatrists interview, Interaction Cronograph, most questionnaires.

Note.—The examples are intended to illustrate a category and are not exhaustive. Those preceding the semicolon are from Buros (1970); those following it are from Cattell and Warburton (1967).

opposed to motivational states that predispose people to see certain kinds of content (see McCall's and Eron's reviews of the Rorschach in PTR). Behind this seemingly measurement oriented controversy there is our very real ignorance of the principles underlying the apperception of "ambiguous" stimuli. Where questionnaires are concerned there has been much controversy about the relative importance of behavioral traits like anxiety or extroversion as opposed to evaluative consistencies and response sets in determining the response to individual items (see Edward's, Frederiksen's, and Stricker's reviews in PTR). Responses to personality ques-

tionnaires presumably depend on consistencies in the person's actual conduct, on processes that register and store information about these consistencies, and on retrieval and evaluative processes that operate when people are asked to describe themselves in a questionnaire. The relative importance of these processes in different people and on different items is surely an appropriate topic for scientific enquiry; very little is really known about the determinants of self descriptive behavior.

An Index to Psychometric Developments

In his review of the Kent-Rosanoff Free Association Test Wiggins noted that,

"Interest in the free association experiment was so wide-spread at the turn of the century that a full account would be almost indistinguishable from a general history of the psychology of that era"

and yet,

". . . no systematic, large scale efforts have been made to develop the instrument as a 'personality test' in the current usage of these words since its inception in 1910."

Research on the scientific problems underlying test-taking behavior is necessary and desirable, but it is not sufficient to guarantee improved measures. It is also necessary to produce an instrument with desirable psychometric properties. PTR may be used to learn what psychometric problems exist in various areas and what progress is being made in resolving them.

Apropos of Wiggins' comment, we note that the Kent-Rosanoff might be considered the first and most primitive member of a family of association instruments (Family 5). Tendler (1930) devised one of the first sentence completion devices in order to improve the association experiment by controlling and systematically varying the anticipatory aspect of the associative reaction and thus aim the test at specific topics and themes (Rapaport, Gill, and Schafer, 1968). Other test authors with similar purposes have produced association tests with multiple choice formats such as the Association Adjustment Inventory and Cattell's T-14. Since these formats are more suited to personality testing than is

the original, unfocused Kent-Rosanoff they may have received the lions share of the development and validation, even though they are less used in purely academic research on associative behavior. Analogous developments seem to occur in other test families. The primitive ancestor of the measures in Family 4 is probably the TAT. While this test has been much used in research on projective processes, it is not conceded to be valid as a personality measure (see Eron and Jensen's reviews in PTR). Reviewers have been uniformly kinder to instruments that are focused on a restricted range of situations or thema such as the Blacky Test, the Rosenzweig PF and the Tomkins-Horn Picture Arrangement Test.

In Family 3 the much researched Rorschach is psychometrically inferior to later instruments such as the Holtzman Ink Blot Test, The Structured Objective Rorschach, and several Cattell measures such as T-327 and T-369, but the later tests have been developed along two quite different lines. Some stress perceptual styles as determinants of the response while other stress content. A similar situation arises in Family 6 where the primitive ancestor is undoubtedly the psychiatric interview. One line of development from the interview led to the Interaction Cronograph which emphasized the timing of the subject's responses to the interviewer, while neglecting response content. This most interesting device is now, unfortunately, out of print. An alternative and more successful procedure retained the content of the interviewer's questions while eliminating the interviewer. This tactic yielded, of course, the questionnaire.

In Family 1 the subject draws a picture and in Family 2 he constructs something from fixed material components. The most primitive versions of these devices ask for single pictures or constructions, but the newer vehicles are asking for a series of pictures (The Eight Card Redrawing Test) or a series of constructions (Maps, The Picture World Test, T-343). Progress here seems to call for a transition from a single item to a multi-item format.

An Index to Validity Studies

While an interest in purely substantive problems can motivate important sorts of background research on personality measurement, the actual labor of instrument development is likely to be

undertaken only when it promises to serve some useful purpose. Its prospects probably depend in part on the type of validity it manages to achieve. Educational tests which aim at content validity will be valued when they sample skills that are much in demand. Personality devices are usually expected to provide construct or criterion validity, and in either case there will be a tendency to judge a measure valuable in proportion as it predicts nontest, real world performances that are conceded to be important by society. These criteria are usually embedded in a larger social matrix and their value to society may be debated in ethical, economic, or political terms. In order to be useful to the tester, however, these criteria must contain a substantial amount of personality variance and they must also possess purely psychometric virtues such as reliability.

PTR can be read as a history of experience with personality scale validation, but in order to use it in this way one must have a method of extracting and organizing the relevant data. The test classification in Table 4 may not help much because the instruments in every family tend to be validated against the same criteria. Paramount among these criteria are indices of adjustment, of job performance, and of successful response to various therapeutic, custodial, or educational treatments. These are classes of criteria, not individual variables, and in organizing them it will be helpful to pick one class and visualize its members in the context of some relevant social process. We will consider the various adjustment criteria as they are encountered by an individual with a severe behavior disorder. As the symptoms of such a disorder develop, they increasingly disturb either the person who exhibits them or significant individuals in that person's environment—family and friends, for example. The person himself may seek psychiatric help or those closest to him may seek it in his behalf. The person's difficulties are then "diagnosed" and he becomes a "patient" with a nosological label. A course of treatment or custodial care is planned and upon its successful completion the former patient may resume something like his old position in society.

Patient oriented criteria. Mental health criteria may be developed at many points in this process. We shall speak of those applicable before referral as "person oriented" and those applicable after

referral as "patient oriented." The most popular patient oriented criteria have been psychiatric diagnoses or ratings, but there are signs that these criteria are falling out of favor. Some diagnostic categories (psychasthenia) have been scrapped even though measures that are supposed to diagnose them continue to thrive. Some categories have displayed suspiciously little inter-rater reliability and others have had their construct validity as diseases seriously challenged (Szasz, 1961). Adjuncts to or substitutes for psychiatric diagnosis now seem to be under development. As Shaffer (Buros, 1970, page 833) points out, inpatient rating schedules (Buros, 1970, pages 54, 68, 121, 157, 169) offer a new approach to the criterion problem. Nurses or other psychiatric staff observe the patient's ward behavior over a period of days or weeks and describe it by endorsing standardized, factor analyzed, descriptive items. Devices of this sort can probably provide accurate, content valid descriptions of some aspects of psychotic behavior. Personality measures might be validated against them and they have the additional advantage of predicting some of the major costs of psychiatric care such as response to chemotherapy, closed versus open ward status, and length of hospitalization.

The decision to seek psychiatric assistance for one's self has been much used as a criterion, particularly with educated middle-class groups. Some questionnaires and sentence completion tests claim sizable coefficients of concurrent validity against this criterion, but this sort of claim has been disparaged by Shaffer (Buros, 1970, page 1232) who points out that voluntary patients tend to exaggerate their symptoms in order to document their pleas for help. The prediction of this criterion has nevertheless been of real interest to universities, military organizations, and other institutions that are committed to providing free health services and who are anxious to minimize costs from this quarter.

Person oriented criteria. It is obviously desirable to develop indices of behavioral disturbance that are applicable to the early stages of the disorder as these are observed in everyday life. Many people will suppose that school teachers are ideally situated to provide criterion descriptions of children with behavior problems because the teachers experience these problems as a cost-producing factor in their own class rooms. Since 1925 at least 30 rating forms for teachers' assessments of pupils have been

created. Half of these are now out of print, however, and none seem to have been extensively used as a criterion for personality test development or indeed (with two or three honorable exceptions) for research of any kind. Classroom behavior ratings present an unusually severe problem for criterion development and a glance at some of the reviews accorded these instruments in PTR will show why this is so.

As Gambrill (Buros, 1970, page 304) points out, some rating schedules are designed "as service instruments to help teachers in understanding pupils rather than for research purposes." If service to the teacher has been emphasized in advertising these instruments, it may be for the reasons mentioned by Lundy (Buros, 1970, page 793). He points out that completing such ratings requires so much clerical time and labor that teachers who have been requested to use them are likely to comply in only the most perfunctory way. This is particularly true when there are no facilities to which obstreperous children can be referred or when parents refuse to accept evaluations of their children that they did not initiate. Some reviewers fear that the use of ratings in the schools will result only in the attachment of labels like "maladjusted" or "emotionally handicapped" to the permanent school records of some of the children and that this will do far more harm than good, especially in view of the naivete about behavioral disorders that prevails among school teachers, administrators, and the general public.

Very little seems to have been done on the development of adjustment criteria in institutions outside the school. The institution most affected by a serious psychiatric disorder is likely to be the family but only one rating form has been developed to record information about the patient's disturbed behavior as seen by other members of his family. This device is very new and has not yet been reviewed (Buros, 1970, page 66). PTR contains almost no information about the development of maladjustment criteria for business organizations. One lone entry concerns a scale on which subordinates rate college administrators or business executives, but these ratings are supposedly used only for the self-development of the person being rated—they are not supposed to be seen or utilized by the ratee's own supervisors.

The neglect of the criterion problem. While personality testers

would benefit from the existence of adjustment criteria reflecting the costs of behavioral disorders in everyday life, the developmental research that is necessary to bring them into being will require the consideration of ethical, sociological, economic, and possibly political problems that transcend the psychometrist's traditional concerns for items, norms, and manuals. Buros' reviewers uniformly underestimate the importance of this work. While a validity coefficient must be evaluated against two reliabilities, most reviewers assess reliability for the test only and make no mention of comparable figures for the criteria. Only rarely do reviewers such as Hanawalt (Buros, 1970, pages 482 and 810) use criterion reliability data to edit validity studies so as to give a clearer picture of the true worth of the instrument. Here and there reviewers may mention the contamination of some criterion rating, but the subtler problems of rating validity are almost never discussed. When ratings are used as criteria, reliability is not enough. There are studies, for example by Freeberg (1967), which show that observer ratings of demonstrable invalidity may be quite homogenous and may show high correlations with other equally invalid observer ratings.

Summary

Although valuable enough as a guide to professional opinion about specific instruments, Buros' *Personality Tests and Reviews* is even more important as an informant on the history and current status of the whole thrust of the personality assessment movement. Its services are available to anyone with a method of asking questions. We conclude that the nonprojective devices (which are chiefly questionnaires) are in a thriving state of growth, but that projective measures are being copyrighted and used in research much less frequently since 1955. New work on these latter devices could quite conceivably cease within the next ten years.

Projective test critics have often called for technical improvement, but Buros' book suggests that the pace of improvement is slow in every area of personality assessment. There is almost no evidence that improvements in projective measures are likely to accrue from competition and the survival of the fittest, from the accumulation of research evidence, or from the stricter application of the current professional standards. Instead of being improved,

projective test ideas that once seemed highly creative may now be in the process of being abandoned.

Since the projective rubric encompasses by far the greater variety of copyrighted personality devices psychologists may resist abandoning them in favor of the questionnaire. They will find PTR full of suggestions for reversing the trend. It is a record of controversies that betray our ignorance about the real processes that mediate test behavior. It is a compendium of descriptive data that may be used to order instruments into developmental series within larger format families. It is a source of information on the problems of validating personality measures against real-world criterion behaviors. An appropriate utilization of these resources might benefit not only personality assessment but the world of personality research, teaching, and service that has come to depend on copyrighted measuring devices so heavily.

REFERENCES

- Buros, O. K. *Personality tests and reviews*. Highland Park, N. J.: Gryphon, 1970.
- Cattell, R. B. *Personality and motivation structure and measurement*. Yonkers-on-Hudson: World Book, 1957.
- Cattell, R. B. and Warburton, F. W. *Objective personality and motivation tests*. Urbana, Ill.: University of Illinois, 1967.
- Cohen, J. The statistical power of abnormal-social research: a review. *Journal of Abnormal and Social Psychology*, 1962, 65, 145-153.
- DuBois, P. H. Varieties of psychological test homogeneity. *American Psychologist*, 1970, 25, 532-536.
- Edwards, A. L. *The measurement of personality traits*. New York: Holt, Rinehart and Winston, 1970.
- Freeberg, N. E. Relevance of rater-ratee acquaintance in the validity and reliability of ratings. Research Bulletin 67-55. Princeton, N. J.: Educational Testing Service, 1967.
- Guilford, J. P. *The nature of human intelligence*. New York: McGraw-Hill, 1967.
- Louittit, C. M. and Browne, C. G. Psychometric instruments in psychological clinics. *Journal of Consulting Psychology*, 1947, 11, 49-54.
- Price, D. J. *Science since Babylon*. New Haven: Yale University, 1961.
- Price, D. J. *Little science, big science*. New York: Columbia University, 1963.
- Rapaport, D., Gill, M. M., and Schafer, R. *Diagnostic psychological testing*, Revised edition. Robert Holt (Ed). New York: International Universities Press, 1968.

- Sundberg, N. The practice of psychological testing in clinical services in the United States. *American Psychologist*, 1961, 79-83.
- Symonds, P. M. *Diagnosing personality and conduct*. New York: Appleton-Century, 1931.
- Szasz, T. S. *The myth of mental illness*. New York, Harper & Row, 1961. *Technical recommendations for psychological tests and diagnostic techniques*. *Psychological Bulletin*, 1954, 51, Supplement.
- Tendler, A. D. A preliminary report on a test for emotional insight. *Journal of Applied Psychology*, 1930, 14, 123-136.
- Woodworth, R. S. and Schlosberg, H. *Experimental psychology*. New York: Holt, 1958.
- Zipf, G. K. *Human behavior and the principle of least effort*. Cambridge: Addison Wesley, 1949.

ELECTRONIC COMPUTER PROGRAM AND ACCOUNTING MACHINE PROCEDURES

WILLIAM B. MICHAEL, Editor
University of Southern California

JOAN J. MICHAEL, Assistant Editor
California State College, Long Beach

<i>Computer Programs for Test Objective and Item Banking.</i> WILLIAM P. GORTH, DWIGHT W. ALLEN, AND ARAM GRAY- SON	245
<i>Interactions Among Group Regressions: Testing Homogeneity of Group Regressions and Plotting Regions of Significance.</i> GARY D. BORICH	251
<i>A FORTRAN Program for the Analysis of Linear Composite Variance.</i> JOHN A. CREAGER	255
<i>Factor Similarity.</i> CARMELO TERRANOVA	261
<i>Eigenvalues and Vectors of Large Matrices on the IBM-1130.</i> JOHN R. HOWELL AND SHARON L. CREWS	263
<i>A Computer Program for Estimating Relative Sequential Con- straint.</i> WILLIAM B. RUDOLPH AND ROBERT B. KANE	267
<i>An Alteration of Program UTEST to Determine the Direction of Group Differences for the Mann-Whitney U Test.</i> STEVEN M. JUNG, DEWEY LIPE, AND THOMAS J. QUIRK	269
<i>A Computer Program for Nonparametric Post Hoc Compari- sons for Trend.</i> JAMES J. ROBERGE	275
<i>A Computer Program for Trend Analysis in a Two- or Three- Factor Experiment with Repeated Measures on One of the Factors.</i> JAMES J. ROBERGE	279
<i>A Streamlined Version of the ALDOUS Simulation of Per- sonality.</i> ROBERT A. LEWIS	283
<i>A Computer Program for Estimating the Power of Tests of Assumptions of Markov Chains.</i> ROBERT W. LISSITZ AND SILAS HALPERIN	287
<i>Subroutine to Decode IBM 1230 Data.</i> JOHN W. MENNE AND JOHN E. KLINGENSMITH	293
<i>Computer Programs for Rank Analysis of Covariance.</i> ED- WARD P. LABINOWICH AND JAMES K. BREWER	295

COMPUTER PROGRAMS FOR TEST OBJECTIVE AND ITEM BANKING¹

WILLIAM P. GORTH AND DWIGHT W. ALLEN

The University of Massachusetts

ARAM GRAYSON

Stanford University

ITEM banking as well as its logical predecessor objective banking are becoming increasingly important in educational measurement. Several projects, both in the U. S. and England, have investigated various forms of item banking. The motivation for these projects is usually the following:

1. To make available to educators better test items for use in examinations in schools;
 2. To provide test items with known item characteristics so that results will be more valid and reliable than those based on locally developed items and that results from one setting can be compared with those from another setting;
 3. To make teachers more familiar, in general, with modern notions of test construction including the classification of test items by categories which they measure, e.g., behavioral objectives;
 4. To utilize test items, which are written by skilled authors, in many contexts without the added costs of writing new ones;
- and

¹ The research herein was performed pursuant to a grant from the Charles F. Kettering Foundation to Dwight W. Allen and under the direction of William P. Gorth, both of the School of Education, The University of Massachusetts, Amherst 01002. Dr. Conrad Wogrin, Director, Research Computing Center, The University of Massachusetts/Amherst, gave the authors the full support and cooperation of his staff and facilities in the final debugging and installation phase of the computer program development.

5. To provide a basis for better decision-making regarding the placement and instructional treatment of students which will minimize losses in students' time and effort.

Objective and item banking require several operations including stocking of the bank, retrieving information from the bank, and using the retrieved information in a variety of testing situations. Each of these operations could be used to characterize existing or future objective and item banks. Stocking the bank consists of writing or compiling objectives and items which have been classified by content and characterized by items statistics. Retrieving from the bank consists of finding the objectives and items which are appropriate to the purpose for which they will be used. Using materials from the bank consists of diagnostic testing, placement testing, criterion-referenced testing within a course, pretesting for the different instructional treatments, or testing on a longitudinal basis using item sampling.

Existing efforts in objective and item banking may be characterized by their purpose and operation. One of the most publicized efforts is the Instructional Objective Exchange, IOX (Popham, 1970). IOX is an attempt to make available to teachers instructional objectives, with a grade level, content, and taxonomical classification of objectives, and sample test items. These materials are made available in the form of mimeographed booklets for specific subject areas and grade levels. The materials are not distributed in a form that can be used directly, i.e., the objectives are not printed in the form that could be transferred to a specific school situation, test items are not numerous enough or appropriately formatted to constitute a test. IOX is not directed toward immediate implementation of objectives or testing programs but more as a guide toward development of locally based objectives and items.

A second major effort would be the Computer-Based Test Development Center, COMBAT (Walter, 1970). The major purpose of COMBAT is to make a large number of teacher written test questions available to classroom teachers for their classroom testing. The classification for test items is by key word and the item statistics are not measured. The storage and retrieval of the items is done by computer. They can be printed immediately in a form which can be used as a classroom test. The computer printing

can be done on masters which duplicate more copies. The testing materials are designed for usual classroom testing.

A more sophisticated bank was produced at the National Foundation for Educational Research in England and Wales because extensive information about item characteristics was available. The work of the Foundation is described by Wood and Skurnik (1969). The item bank includes items which can be used by school-based examiners to determine the score of students in the certification of secondary education in England and Wales in mathematics. Extensive work went into the development of the item bank. Items were classified by task. They were pretested so that their item characteristics were known. The storage and retrieval of items was from a card file.

Another effort in objective and item banking is the focus of this paper. The banking system has been developed by the Project for Comprehensive Achievement Monitoring, CAM (Gorth, 1968). CAM has developed a model of evaluation useful in curriculum evaluation and classroom management. The model consists of longitudinal testing, using item sampling, of the specific behavioral objectives for a course. In order to support the testing activity, which uses a large number of test forms, and therefore, a large number of test items, computer programs were developed to streamline test development.

All of the items in the CAM item bank have been classified in at least three dimensions: their content, their taxonomical level, and the sequence in which they are taught in the typical school course. The relationships between items and objectives are referenced and item analysis information at pretest, posttest, and retention time intervals is available. Both the objectives and the test items are stored on magnetic tape. The classification of items and their relation to the objective is also recorded in the computerized item bank. The objectives and items may be selected on a preliminary basis for perusal by individual teachers or for establishing a bank consisting of a subset of the total objectives and items. After final selection and allocation to test forms, the tests can be printed in an easily duplicatable format. The answer keys for the tests are both printed as well as punched into computer cards for quick analysis by other programs available from CAM (Gorth, Grayson, and Lindeman, 1969; Gorth, Grayson, and Stroud, 1969; Gorth, Grayson, Popejoy, and Stroud, 1969).

*The CAM Computer Programs**Subsystem 1: Storing and Editing of Objectives and Items*

Input. The input into the computer program for storing items and objectives consists of three different types of data cards. The first type of data card contains the identification number of the objective, its classification as to subject and grade level, and its text. The text is keypunched onto computer cards in the format in which it will be later printed. The second type of data cards consists of test items keypunched onto computer cards with the text of the test question formatted so that it may be printed directly. Each question may have as many as nine alternatives, and the correct alternative can be indicated. Each question is also permitted to have explanatory notes to the examiner which specify certain diagrams, maps, figures, or physical objects which the student will be given at the time of the test. The third type of data cards consists of the classification scheme for the test item. Each test item is classified by subject and grade level as well as by other content and psychological classifications and item statistics. The classification information is limited to 15 distinct categories. Modification or additions may be made to the classifiers and the text of the item after the item bank has been initially developed. The capability of modifying the classifiers would allow additional item analysis information to be added after the item bank was created. The item bank resides on magnetic computer tape and additional objectives and items can be added or deleted as the occasion requires.

Output. The output from the computer program is a magnetic tape and its printed listing. A series of error messages is available to inform the user of obvious errors in the processing of data. Data, which are in error, will not be recorded in the bank. Thus, possible inappropriate information in the item bank is eliminated.

Subsystem 2: Preliminary Selection of Objectives and Items

Input. The second computer program selects test objectives and their associated test items by the classifiers stored with the test items. The items are chosen by an internal compiler which allows specific values of any classifier or any combination of classifiers, e.g., content area, taxonomic level, and item characteristics, to be used as selection criteria for items and objectives. The desired

criteria are read into the computer program. The program reads the data bank tape and selects the items by the criteria specified.

Output. The output from this stage of the item banking program consists of two parts. The first part is a printout of the objectives on one side of a page of computer output and of the associated test items on the other side. If there are more items than one for an objective, they are all printed in the same part of the output one after another. The objective is printed only once next to the first item in the list of items associated with it. Each item is printed out in the format which would be used on a test. The correct answer to an item is indicated when it is printed out. Its identification number in the item bank is also printed out for later specification.

The second form of output is a magnetic tape which is a summary of the items selected for this set of tests at the first stage. These items are recorded on a second data tape so that they will not have to be relocated in the master objective and item bank. If the bank increases in size to the level of fifty or one hundred thousand test items, a preselection or preevaluation of the appropriateness of items by teachers must be made, but the expense of searching the data tape for the preliminary selection of items should not be duplicated. Therefore, the subset of items and objectives is recorded on the intermediate tape which would be much shorter than the original tape and tailored to the specific needs of the teacher. The subset may also serve as an objective and item bank tailored to local needs and can be distributed for local use.

Subsystem 3: Objective and Test Printing

Input. The magnetic computer tape (containing the items and objectives selected by individual teachers for specific testing situations) which was written at the preceding stage in the item banking sequence is used as input for the final stage in the item banking sequence. Also the final selection and arrangement of items on test forms are specified on data cards.

Output. The output from the third stage of the item banking procedure consists of tests printed in a form which can be immediately duplicated and administered to examinees. The tests may be printed on duplication masters or directly on multi-part, computer output paper. Each page of the test is labeled by the school, the course, the test form, and the page number.

In addition to the tests which are printed in a form which can be used directly, each of the objectives associated with the test is printed out in a form which also can be directly duplicated and distributed to examinees or students in a course. The objectives are numbered according to the pattern and sequence which has been chosen by the teacher to fit his curriculum organization. An answer key for each test is also provided. The answer key is printed and punched in a format which is appropriate for the analysis programs developed by Project CAM (Gorth, 1968). Each answer key contains all the information concerning the classification of each item on the test. The alternatives to the test items are randomized by the computer program before they are printed out in the final version of the test.

Additional information about the programs and their availability may be obtained from William Gorth, Director, Project CAM, School of Education, The University of Massachusetts, Amherst, Massachusetts 01002.

REFERENCES

- Gorth, W. P. (Organizer) Comprehensive achievement monitoring. Symposium presented at the Annual Meeting of the American Educational Research Association in Los Angeles, 1969.
- Gorth, W., Grayson, A., and Lindeman, R. A computer program to evaluate item performance by internal and external criteria in a longitudinal testing program using item sampling. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 181-183.
- Gorth, W., Grayson, A., and Stroud, T. A computer program to tabulate and plot achievement profiles of longitudinal achievement testing using item sampling. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 179-180.
- Gorth, W., Grayson, A., Popejoy, L., and Stroud, T. A tape-based data bank from educational research or instructional testing using longitudinal item sampling. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 175-177.
- Popham, J. W. The Instructional Objectives Exchange: Progress and Prospects, Paper presented at the Annual Meeting of the American Educational Research Association in Minneapolis, March 1970.
- Walter, L. J. COMBAT: A system now in operation. Paper presented at the Annual Meeting of the National Council on Measurement in Education in Minneapolis, March 1970.
- Wood, R. and Skurnik, L. S. *Item banking*. London, England: National Foundation for Educational Research in England and Wales, 1969.

INTERACTIONS AMONG GROUP REGRESSIONS: TESTING HOMOGENEITY OF GROUP REGRESSIONS AND PLOTTING REGIONS OF SIGNIFICANCE

GARY D. BORICH

Institute for Child Study
Indiana University

RECENT interest in aptitude-treatment research has revitalized procedures for determining interactions among group regressions. Among these procedures is the homogeneity of group regressions test and the Johnson-Neyman technique. The homogeneity of group regressions model (Walker and Lev, 1953; Edwards, 1968) tests the hypothesis that regression slopes are equal across treatments, while the Johnson-Neyman technique (1936) determines regions of significance and nonsignificance when the equal slopes hypothesis is rejected.

Two programs have been reported which analyze aptitude-treatment interactions; both, however, are limited in scope. Ter-ranova (1970) has programed an F -test for the homogeneity of group regressions, but because group interactions were not of primary interest, the program does not determine regions of significance. A second programming effort by Carroll and Wilson (1970) has produced a program which determines regions of significance without first testing for homogeneity of group regressions. Their program determines regions of significance for the case in which there are two groups and two predictor variables but does not plot the data along the regression lines to indicate where such regions are meaningful. Although one option to the Carroll and Wilson program is that data input may be in the form of means, standard deviations, and correlations, it is important to note that the Johnson-Neyman technique assumes linearity of regression as well

as significant correlations between predictor and criterion. Therefore, correlations and scatterplots should be computed before the Carroll and Wilson program is used.

Program Description

The program combines the essentials of the Terranova and the Carroll and Wilson programs for the case in which there are two groups, one predictor variable, and one criterion variable. The program plots data points, regression lines, and region(s) of significance. As suggested by Abelson (1953), the homogeneity of regression slopes test is performed, after which regions of significance are determined, if applicable.

Input to the program consists of (a) a selected probability level for the Johnson-Neyman test, (b) slope and constant for the treatment groups, (c) format for each group, (d) N 's for each group, and (e) data cards, with criterion first and predictor second.

The program computes and prints:

1. F -test for homogeneity of regressions, and if applicable:
2. Point at which regression lines intersect.
3. Lower boundary of significance.
4. Upper boundary of significance.

The program plots:

1. Scattergram for two treatments with data points for each treatment coded differently
2. Regression lines among data points.
3. Region(s) of significance with boundary points plotted through data.

The abscissa is automatically scaled to include all aptitude values, and the ordinate is automatically scaled with predicted and obtained criterion values. Sample problem solutions and plots are available from the author

Summary

The Johnson-Neyman technique assumes that regression lines for each treatment are linear and slopes unequal. A scatterplot for the linearity of regression lines assumption and an F -test for the more restrictive assumption of homogeneity of regression slopes

are provided by the program with a plot of the region(s) of significance. The program provides all necessary calculations for the aptitude-treatment investigation in which there are two groups, one aptitude, and one criterion.

REFERENCES

- Abelson, R. P. A note on the Neyman-Johnson technique. *Psychometrika*, 1953, 18, 213-218.
- Carroll, J. B. and Wilson, G. F. An interactive-computer program for the Johnson-Neyman technique in the case of two groups, two predictor variables, and one criterion variable. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1970, 30, 121-132.
- Edwards, A. L. *Experimental design in psychological research* (3rd ed.), New York: Holt, 1968.
- Johnson, P. O. and Neyman, J. Tests of certain linear hypotheses and their applications to some educational problems. *Statistical Research Memoirs*, 1936, 1, 57-63.
- Terranova, C. *F-test for the homogeneity of regression assumption*
- Walker, H. and Lev, J. *Statistical inference*. New York: Holt, 1953.
- EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1970, 30, 155-156.
- of the analysis of covariance model. *EDUCATIONAL AND PSYCHO-*

A FORTRAN PROGRAM FOR THE ANALYSIS OF LINEAR COMPOSITE VARIANCE

JOHN A. CREAGER

American Council on Education

WHEN regression, discriminant, and canonical models are applied to empirical data, the result is one or more linear composites of variables. Interest in such composites seldom lies solely in maximizing prediction, but also in some insight into the interrelations among the variables in their roles of contributing to that prediction. Typically, the resulting linear composites are examined for what is involved algebraically in the prediction system, and for what is involved empirically in the research context in which the data were obtained. Both informal procedures, e.g., "eyeballing the weights," and more sophisticated techniques, involving the terms of the standard formula for linear composite variance, partial or part correlations, or analysis of covariance, are unsatisfactory because they fail to cope adequately with the multicollinearity of the system. In order to cope with multicollinearity in the analysis of a prediction system, it is necessary "to grasp the nettle" and analyze the role of the multicollinearity, itself. Otherwise, we shall continue to be confounded in our judgments of the relative import of the variables defining prediction composites.

A procedure for accomplishing such analysis of linear composites has been proposed, and has been illustrated for the special case of regression composites, by Creager and Boruch (1969). Subsequent development has shown the procedure to be completely general and therefore applicable to the analysis of left and right composites associated with one or more canonical roots. The basis of the procedure is the determination of loadings for the linear composites on factors defined by complete orthogonal factor analy-

sis of the correlation matrix used to develop the composites. Squared loadings for the composites are completely independent and additive portions of the total linear composite variance. To be of practical value, the orthogonal factors must have substantive meaning.

Program COMPVAR implements the Creager-Boruch procedure for orthogonal analysis of linear composite variance, given the results of a complete orthogonal factor analysis of a prediction system. Variable formatting permits one to take the punched output from a factor program, e.g., Restricted Maximum Likelihood Factor Analysis (Joreskog and Gruvaeus, 1967) or VARIMAX (Kaiser, 1959) and obtain loadings on one to six linear composites, based on the set of variables factored. This flexibility permits analysis of several regression composites using the same predictors but varying criteria, up to six discriminant functions, or the left and right composites from three canonicals. Composites are assumed to be defined in terms of *standard* scores in the components.

The program is written in FORTRAN IV, and includes two subroutines: MATMLY (matrix multiplication) and RESYMA, which reads in and prints the correlation matrix. I/O is entirely card input and on-line, printed output.

Input

Input will be described in terms of the data deck structure.

1. Problem Card in 20A4 permits user to label his run.
2. Parameter Card in 2I2,I1:
 - NVAR, the number of variables, up to 50.
 - NCOM, the number of common factors, up to 50.
 - NLIN, the number of composites, up to 6.
3. Two variable format cards in 20A4, the first designating format for common factor loadings, the second designating format for vectors of *squared* uniqueness loadings and of composite weights.
4. A third variable format card in 18A4 designating format for reading in the *lower triangular* correlation matrix in subroutine RESYMA.
5. The correlation matrix. Each row is punched from the left up to, but excluding the diagonal, the rest of the card being left blank. Where the number of variables and format require, a slash

may be used in the variable format and continuation cards used for each row of the matrix.

6. Cards containing factor loadings. There will be one set of cards for each row (i.e., variable of the factor matrix). Each set of cards will contain the *common* factor loadings for that variable.

7. Sufficient cards in format to read in a $1 \times \text{NVAR}$ vector of squared uniqueness loadings. The program converts this vector to diagonal matrix form.

8. Cards containing the composite weights. There will be one card (or set of cards) for each composite, reading in a vector ($1 \times \text{NVAR}$) of weights.

Output

The print-out exhibits the following:

1. Contents of the problem card, parameter card, and three variable format cards read in.
2. The input correlation matrix in full symmetric form with unit diagonals.
3. The common factor matrix in usual form.
4. A column vector of squared uniqueness loadings.
5. The transpose ($\text{NVAR} \times \text{NLIN}$) of the weights matrix. Thus, the weights for a given composite are found as a column vector in this matrix.
6. The variance-covariance matrix of raw score composites of standard scores, WRW' .
7. A column vector of composite standard deviations, the square root of diagonal values in the previous matrix. These values are explicitly required for computations in the Creager-Boruch procedure.
8. The $\text{NLIN} \times \text{NCOM}$ matrix of estimated loadings for composites on the common factors. For a given composite, the loadings are found as a row vector in this matrix.
9. The $\text{NVAR} \times \text{NLIN}$ matrix of estimated loadings for composites on the unique factors. These values are *not* squared. For a given composite, the loadings are found as a column vector in this matrix.
10. The variance-covariance matrix of standard score composites (i.e., composite intercorrelation matrix) developed from multiplying factor loadings (both common and unique) for composites by

its transpose. Diagonal values may fall short of unity by an amount dependent upon the completeness of the factoring of the original correlation matrix. The square root of these values is computed and printed as a column vector of standard deviations of these composites.

11. The $NLIN \times NCOM$ matrix of *squared* loadings for composites on common factors.

12. The $NVAR \times NLIN$ matrix of *squared* loadings for composites on unique factors.

Comments

Program COMPVAR has been debugged on the XDS Sigma Five computer, using the hypothetical example problem in the original Creager-Boruch paper, and with real data from an achievement study (Jones, 1963).¹ The latter involved four predictors and two criteria, and developed two canonicals (four composites) and multiples for each criterion. All six composites were analyzed in a single pass of the program requiring 1.17 minutes including compilation. Actual execution time was .2 minute. Liberal use of comment cards and judicious selection of program parameter names make the program easy to read and follow.

The exhaustive factoring of the correlation matrix is critical for complete account of system variance. One should, in fact, define composites using the same correlation matrix as the one actually factored, i.e., actually *reproduced* by the factor solution. If one is using an estimation procedure like maximum likelihood for factoring, he will get the population estimate of the correlation matrix and can use this in the production and analysis of composites (e.g., regression, discriminant, or canonical). If one is working entirely from the sample data, an algebraic factor solution such as VARIMAX will give a reasonable solution, provided sufficient principal components have been rotated to obtain a clean separation of common and unique factors. Serious consideration should also be given to prior correction of the correlation matrix for attenuation.

Availability

Copies of the source program deck, sample print-out, and program documentation may be obtained for the nominal cost of reproduction

¹ Dr. Jones kindly supplied computer printout of his canonical analysis.

and mailing. Requests should be addressed to the author, Office of Research, American Council on Education, One Dupont Circle, Washington, D.C. 20036.

REFERENCES

- Creager, J. A. and Boruch, R. F. Orthogonal analysis of linear composite variance. *Proceedings of the American Psychological Association*, 1969, 113-114.
- Jones, K. J. Predicting achievement in chemistry: A model. *Journal of Research in Science Teaching*, 1963, 1, 226-231.
- Joreskog, K. G. and Gruvaeus, G. RMLFA, a computer program for restricted maximum likelihood factor analysis. (Research Bulletin RB-67-21) Princeton, N. J.: Educational Testing Service, 1967.
- Kaiser, H. F. Computer program for VARIMAX rotation in factor analysis. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1959, 19, 412-420.

THE JOURNAL OF THE
ROYAL ANTHROPOLOGICAL INSTITUTE
OF GREAT BRITAIN AND IRELAND
VOLUME 111, PART 1, 1981
PAGES 1-114
CONTENTS
The Journal of the Royal Anthropological Institute of Great Britain and Ireland, Volume 111, Part 1, 1981, contains 114 pages of research and review articles. The articles are organized into sections: 'Human Evolution', 'Human Variation', 'Human Development', 'Human Health', 'Human Behaviour', 'Human Society', 'Human Culture', 'Human Language', 'Human Thought', 'Human Emotion', 'Human Perception', 'Human Memory', 'Human Learning', 'Human Reasoning', 'Human Creativity', 'Human Innovation', 'Human Progress', 'Human Future'. The articles are written by leading experts in their fields and provide a comprehensive overview of the current state of research in human anthropology. The journal is published by the Royal Anthropological Institute of Great Britain and Ireland, which is a leading international organization for the study of human evolution, variation, and development. The journal is available in both print and electronic formats, and is a valuable resource for researchers and students in the field of human anthropology.

FACTOR SIMILARITY

CARMELO TERRANOVA

Educational Research Council of America, Cleveland

AN intuitively reasonable measure of attitude change, when using a semantic differential instrument, is the degree of incongruity between the factors extracted from similar concepts of pre- and post-administrations. The wider use of the semantic differential as a measurement instrument for attitude change has been somewhat restricted by the unwieldy amount of data generated and by the relative inaccessibility of an incongruity measure. Harmon (1960) and Tucker, Koopman, and Linn (1969) described a coefficient of congruence that measures the degree of similarity between a factor of one matrix and a factor of another matrix.

Description of Program

The program was designed to compare factors from similar concept matrices (person by scale) obtained at different times. It also may be used to compare factors of different concepts. It allows the utilization of all variables (scales) in determining congruence between factors rather than by comparing factors by means of a subset of variables (i.e., activity, potency, or evaluation).

Input consists of (a) the number of concept pairs being compared, (b) the identification numbers of the concepts, (c) the number of factors in each concept, (d) the number of variables comprising each factor, and (e) the factor loadings of the previously rotated solution. The following is computed and printed for each pair of factors: (a) the sum over the variables of the products of the factor loadings, and (b) the coefficient of congruence.

Discussion

The usefulness of the semantic differential as a measurement instrument for attitude change may be increased by this procedure—

that is, by computing indices of factor similarity and subsequently interpreting the degree of dissimilarity as evidence of attitude change; conversely, the degree of similarity or congruence may be indicative of the stability of attitudes. The computation of the coefficient of congruence provided by this program ought to aid those investigators who would like to use this intuitively pleasing measure of attitude change.

REFERENCES

- Harmon, H. H. *Modern factor analysis*. Chicago: The University of Chicago Press, 1960.
- Tucker, L. R., Koopman, R. F., and Linn, R. L. Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, 1969, 34, 421-459.

EIGENVALUES AND VECTORS OF LARGE MATRICES ON THE IBM-1130

JOHN R. HOWELL AND SHARON L. CREWS
Virginia Commonwealth University

This FORTRAN language program calculates the eigenvalues and eigenvectors of (or "diagonalizes") a real symmetric $N \times N$ matrix by Jacobi's method as described by Ralston (1965). A slow version (3.6 microsecond) IBM-1130 computer with disk and 16K words of core storage was used. Eigenvalues and vectors, of course, are required in a number of multivariate statistical methods including principal components and factor analysis.

A straightforward application of the Jacobi method with the eigenvalues and vectors being developed in two square arrays in core revealed that a matrix up to size $N = 50$ could be handled. The computer time required for $N(N - 1)/2$ plane rotations for both arrays (for $N = 50$) was 20 minutes.

Diagonalization of Larger Matrices Using Disk

It was felt that one should be able to use the large storage capacity of disk in order to diagonalize matrices larger than size 50.

Assuming that a real, symmetric, $N \times N$ matrix to be diagonalized is stored in File 1 on the disk and an $N \times N$ identity matrix is stored there on File 2, one proceeds as follows:

1. First, read the matrix from File 1 into core and perform $N(N - 1)/2$ plane rotations. Store the $N(N - 1)/2$ sine and cosine pairs used for these rotations on File 3 on disk.
2. Now store this partially diagonalized matrix back on File 1 and move the identity matrix (later the partially developed eigenvectors) from File 2 on disk to the same array in core that was used in step 1.

3. Using the sine and cosine pairs that were just stored on File 3 on disk, perform the same $N(N - 1)/2$ plane rotations on the identity matrix that were performed on the matrix being diagonalized.
4. Now move these partially developed eigenvectors to File 2 on disk and repeat the entire process (which may be called one cycle) a specified number of times usually about six. If further iterations are required for greater accuracy the program can be executed again and the process continued a specified number of times. This can be done because the partially developed eigenvalues and vectors are stored permanently in disk files.

Tactics

The table below shows the IBM-1130 computer times required for one cycle for matrices from size 50 to size 80, the largest that can be handled by this approach on an IBM-1130 computer (with 16K words of core storage). Any given matrix between these sizes can be stored on disk during any working day. As soon as a large block of computing time is available, such as nights, weekends, or holidays, the program can be executed, using the above procedure. The computer can be left unattended during this time, since there is minimal print-out (and thus, little danger of a paper jam) until the end of computations. The average absolute value of this sine for the $N(N - 1)/2$ plane rotations is printed. This value should approach zero as diagonalization is neared. Finally, the diagonalized matrix of eigenvalues and the matrix of eigenvectors are printed.

Use of Results

When one is satisfied that the required eigenvalues and vectors have been obtained with sufficient accuracy, any existing main pro-

TABLE 1
Cycle Times for Matrix Sizes

Matrix Size	Cycle Time (Minutes)
50	47
60	69
70	101
80	151

Note.—These points plot as a straight line on semi-logarithmic paper.

gram that calls a matrix diagonalizing subroutine can be modified to read the required numbers from disk.

Availability

Program listings can be obtained by writing to John R. Howell, Department of Biometry, Virginia Commonwealth University, Box 832, MCV Station, Richmond, Virginia 23219.

REFERENCE

Ralston, Anthony. *A first course in numerical analysis*. New York: McGraw-Hill, 1965.

A COMPUTER PROGRAM FOR ESTIMATING RELATIVE SEQUENTIAL CONSTRAINT

WILLIAM B. RUDOLPH

Iowa State University

ROBERT B. KANE

Purdue University

THE inception of information theory (Shannon, 1948; Wiener, 1948) gave researchers an additional tool to study language. However, the ubiquitous enigma in the application of information theory concepts to language analysis is the determination of entropy and the subsequent redundancy. Garner and Carson (1960) separated redundancy into two parts; distributional constraint and sequential constraint. Binder and Wolin (1964) proved sequential constraint was equal to the multiple contingent uncertainty. A model formulated by Newman and Gerstman (1952) may be adapted to estimate the multiple contingent uncertainty and, consequently, the relative sequential constraint. Briefly, the constraint imposed on the criterion variable (symbol being predicted) by each of the predictor variables (preceding m symbols where m is a positive integer) is determined. These constraints are then summed resulting in an estimate of sequential constraint.

Carterette and Jones (1963) computed relative sequential constraints for a variety of childrens books as well as for biblical passages and adult literature, all written in English. Their program accommodated 28 distinct characters (26 letters, end of word, end of sentence).

The study of relative sequential constraints of technical English requires a program which can accommodate many more than 28 distinct characters because of the heavy use of nonalphabetic symbols in technical discourse.

The Program

A program to compute relative sequential constraint for passages containing up to 126 distinct characters was created. Textual material is first encoded into machine characters. Then contingency tables are constructed which show the frequency with which each symbol follows every other symbol immediately and at distances of 2, 3, 4, ..., 120 characters. Finally computations resulting in estimates of the relative sequential constraint are executed.

The user may direct the machine to cut off contingency table construction at any stage between 2 and 120 that he desires. Used on a CDC 6500 computer the program is quite efficient. To illustrate, a 5,000 symbol passage containing 51 distinct characters was analyzed in approximately 16 seconds whereas a 113,097 symbol passage containing 78 distinct characters was analyzed in approximately 367 seconds. These analyses computed sequential constraint for characters separated by a maximum distance of 16.

REFERENCES

- Binder, A. and Wolin, B. R. Informational models and their uses. *Psychometrika*, 1964, 29, 29-54.
- Carterette, E. C. and Jones, M. H. Redundancy in children's texts. *Science*, 1963, 140, 1309-1311.
- Garner, W. R. and Carson, D. H. A multivariate solution of the redundancy of printed English. *Psychological Reports*, 1960, 6, 123-141.
- Newman, E. G. and Gerstman, L. J. A new method for analyzing printed English. *Journal of Experimental Psychology*, 1952, 44, 114-125.
- Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal*, 1948, 27, 379-423.
- Wiener, N. *Cybernetics or control and communication in the animal and the machine*. New York: Wiley, 1948.

AN ALTERATION OF PROGRAM U TEST TO DETERMINE THE DIRECTION OF GROUP DIFFERENCES FOR THE MANN-WHITNEY U TEST

STEVEN M. JUNG AND DEWEY LIPE
American Institutes for Research, Palo Alto

THOMAS J. QUIRK
Educational Testing Service, Princeton

ONE of the most useful, and most used, of nonparametric procedures is the Mann-Whitney U test (Siegel, 1956; Hays, 1963). The U test is used as an alternative to the parametric t test in assigning a probability statement to the differences between two independent samples. Probability in this case is a function of the "smallness" of U , and significance is assigned on the basis of U values which are equal to or less than tabled values found in Siegel. Alternatively, in cases where sample group sizes are large, a z transformation may be made, allowing direct probability estimation from the unit normal distribution.

The calculations necessary to compute U and associated z values are simplified considerably by the use of three FORTRAN subroutines described in the IBM Scientific Subroutine Package (IBM, undated). These subroutines (RANK, TIE, and UTEST), when called by a suitable user-constructed main program, perform the operations of ranking observations in the two sample groups, calculating the sums of ranks corrected for ties, and computing the values of U and z (for suitably large samples). The user can employ his own ingenuity in the construction of his main program, but all that are required are the basic I/O and subroutine call statements.

A deceptive and subtle problem in this procedure is that of determining the direction of obtained differences. One common practice

is to incorporate into the body of the main Mann-Whitney U program some routine for calculating descriptive statistics on the dependent variables for the two sample groups. Or, alternatively, the user may run his data through some available descriptive statistics program prior to the application of the U test program. In either case, the output usually contains such statistics as the mean or median of the dependent measures for each sample group. It is fallacious, however, to assume that these statistics are adequate to indicate the direction of obtained sample differences, especially in situations which have called for the use of a nonparametric test.

In Table 1 a set of contrived data illustrates a situation in which a measure of central tendency shows Group 2 to be the "larger" of two groups, whereas a measure of the relative rank ordering of the scores show Group 1 to be greater.

TABLE 1
Mean and Rank Order Values of Two Sets of Contrived Data

Group 1 ($N = 5$)		Group 2 ($N = 6$)	
Scores	Ranks	Scores	Ranks
0	1	49	4.5
50	9.5	49	4.5
50	9.5	49	4.5
50	9.5	49	4.5
50	9.5	49	4.5
		49	4.5
$\Sigma X_1 = 200$ $\bar{X}_1 = 40$		$\Sigma X_2 = 294$ $\bar{X}_2 = 49$	
$\Sigma R_1 = 39$		$\Sigma R_2 = 27$	

This situation is obviously more likely to occur when one of the two distributions being compared is highly skewed. Yet it is just such a situation which calls for the application of the non-parametric Mann-Whitney U test in preference to a parametric test such as t , which makes use of mean differences. Siegel's (1956) exposition, while lucid in other details, sheds little light directly on the problem of determining the direction of obtained differences. Intuitively and computationally, however, the solution is straightforward.

When applying the sum of ranks method for calculating U , two formulas may be used:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \quad (1)$$

or, equivalently:

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 \quad (2)$$

where

n_1 = number of cases in sample group 1.

n_2 = number of cases in sample group 2.

R_1 = sum of ranks assigned to group 1.

R_2 = sum of ranks assigned to group 2.

and where the lowest score is given a rank of 1 and the highest is given a rank of $n_1 + n_2$.

Formulas (1) and (2) generally yield different values; however, only the smaller of these values is called U and represents the basis for the tables found in Siegel (1956). The larger value is called U' . If only the first formula is used, the transformation

$$U_2 = n_1 n_2 - U_1 \quad (3)$$

may be used to determine the value of U . As it happens, the latter is the procedure which is followed in the Scientific Subroutine Package (SSP) subroutine UTEST. UTEST always returns to the main program the value U which is the smaller of two possible values, U_1 and U_2 . This is correct procedure. However, the investigator is forced to use other data to determine which group produced this U value and was, hence, stochastically the larger.¹

An easy modification may be made to the UTEST subroutine in order to return to the main program values which can be used to determine the direction of group differences. This is to alter the subroutine so that both U_1 and U_2 are returned, enabling the investigator to determine by observation which value is smaller and, hence, which group has the larger sum of ranks.

¹ Siegel confuses this issue in his example (Siegel, 1956, pp. 121-123) by using the value U' to compute a z , which led him to reject the null hypothesis in favor of his group 2. The actual value of U in this case, computed by using formula 2, is 64.0, which yields a similar, although negative, z of -3.45. Since either U or U' will produce identical absolute values of z , it is necessary for the experimenter to know which U value is used and from which formula, 1 or 2, it is derived in order to know which group is larger. There seems to be much less chance of confusion if U is always used rather than U' , since U always represents the stochastically larger group.

In the example presented in Table 1 the U_1 value (associated with group 1) is six and the U_2 value (associated with group 2) is 24. The smaller value, U_1 , becomes U . When compared against the tabled values of U in Siegel (1956, p. 271), this allows the assignment of a .063 probability that so small a value could have occurred by chance. Directionality is determined easily, since the group which produced U is the larger. In the example case, this is group 1.

It may rarely occur, but it certainly is not improbable that, for some studies, U' will equal U . When U' equals U , this means that all the scores in both samples are equal and, therefore, all receive the same rank order.

The characteristics of a program which makes use of the UTEST subroutine modified in this manner are now described.

Input

1. System cards.
2. Header card, to be printed on output.
3. Control card, containing: (a) number of subjects in smaller group; (b) number of subjects in larger group; (c) number of variables; (d) indicator for omission of correction for ties; and (e) indicator for descriptive statistics desired.
4. F -type variable format card.
5. Data cards, with data for smaller group placed first.
6. Finish cards after last data set.

Output

1. Printed identification information as punched on header card above.
2. Descriptive statistics for all variables in smaller group and larger group, in that order.
3. Value of U_1 and U_2 for each variable and value of z computed from the smaller of these values (U) if the number of cases in the larger group is more than 20.

Limitations

The sum of $n_1 + n_2$ cannot exceed 200; and the number of variables cannot exceed 50.

Copies of this program are available upon request from the senior author.

REFERENCES

- Hays, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.
- International Business Machines Corporation, *System/360 scientific subroutine package*, Programmer's Manual # 360A-CM-03X, undated.
- Siegel, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

A COMPUTER PROGRAM FOR NONPARAMETRIC POST HOC COMPARISONS FOR TREND¹

JAMES J. ROBERGE

Temple University

In behavioral science experiments involving K treatments the researcher is usually not satisfied with merely refuting the equality of the treatment effects. Instead, he is often more interested in performing particular post hoc comparisons, e.g., trend comparisons, among the treatments. For parametric analysis of variance models, a common procedure, following the rejection of the null hypothesis of equal expected values, is the use of orthogonal polynomials to estimate the magnitude of these trend comparisons. Recently, Marascuilo and McSweeney (1967) discussed analogous post hoc trend analysis procedures which may be employed following the rejection of the null hypothesis by a nonparametric test such as the Kruskal-Wallis (1952) one-way analysis of variance for rank data, the Friedman (1937) two-way analysis of variance for rank data, or the Cochran (1950) two-way analysis of variance for dichotomous data.

The program described in this paper is designed to perform the aforementioned nonparametric analyses. Specifically, it (a) calculates the statistic for a given nonparametric test, (b) compares this statistic with the chi-square value required for significance at the 5 per cent level, and (c) calculates post hoc confidence intervals for the trend comparisons, if the null hypothesis can be rejected at this level.

Formulas

The formulas presented below are similar to those discussed by Marascuilo and McSweeney (1967). The formulas used to calculate

¹ The author gratefully acknowledges the support for this research which was provided by a Faculty Research grant funded by Temple University.

the statistics H , χ^2 , and Q , for the Kruskal-Wallis, Friedman, and Cochran tests, respectively, are those presented in statistics textbooks commonly used in the behavioral sciences. In the case of the Kruskal-Wallis and Friedman tests, these statistics are corrected for tied ranks.

Contrasts

The arbitrary comparisons, $\hat{\psi}$, of the average ranks (or condition means) are calculated by the formula $\hat{\psi} = a_1 \bar{R}_1 + a_2 \bar{R}_2 + \dots + a_K \bar{R}_K$ where a_1, a_2, \dots, a_K are a set of coefficients of orthogonal polynomials and $\bar{R}_1, \bar{R}_2, \dots, \bar{R}_K$ are the average ranks (or condition means).

Variances

The variances of the various comparisons are calculated by the following formulas:

$$\text{Var}(\hat{\psi}) = \left[\frac{N(N+1)}{12} - \frac{\sum_{s=1}^d (t_s^3 - t_s)}{12(N-1)} \right] \sum_{k=1}^K \frac{a_k^2}{n} \quad (\text{Kruskal-Wallis Model})$$

where N is the total number of observations (or ranks), d is the number of sets of tied observations, t is the number of tied observations for a given set s , K is the number of samples, n is the number of subjects in each sample, and a is as defined above.

$$\text{Var}(\hat{\psi}) = \left[\frac{K(K+1)}{12} - \frac{\sum_{s=1}^d (t_s^3 - t_s)}{12n(K-1)} \right] \sum_{k=1}^K \frac{a_k^2}{n} \quad (\text{Friedman Model})$$

where K is the number of experimental conditions, n is the number of subjects tested under the K conditions, and d , t , and a are as defined above.

$$\text{Var}(\hat{\psi}) = \frac{\left(K \sum_{i=1}^n S_i - \sum_{i=1}^n S_i^2 \right)}{nK(K-1)} \sum_{k=1}^K \frac{a_k^2}{n} \quad (\text{Cochran Model})$$

where S is the number of successes for subject i across the K conditions, and n and a are as defined above.

Confidence Intervals

The confidence intervals for the various comparisons are of the following form:

$$\hat{\psi} - \sqrt{\chi_{K-1}^2 (1-\alpha)} \sqrt{\text{Var}(\hat{\psi})} < \psi < \hat{\psi} + \sqrt{\chi_{K-1}^2 (1-\alpha)} \sqrt{\text{Var}(\hat{\psi})}$$

where χ^2 has $K - 1$ degrees of freedom, $\alpha = .05$, and $\hat{\psi}$ and $\text{Var}(\hat{\psi})$ are as defined above.

Input

The job deck set-up for each analysis is as follows:

Problem card

- Columns 1-2 = number of samples or experimental conditions (K)
3-5 = number of subjects per sample or experimental condition (equal n s are required)
6 = nonparametric test (1 = Kruskal-Wallis;
2 = Friedman; 3 = Cochran)
7 = trend analysis (1 = yes; 0 = no)

Coefficient cards

If the user chooses to have a post hoc trend analysis performed, then these cards contain the linear, quadratic, and cubic (if $K > 3$) coefficients of the orthogonal polynomials; otherwise, they are omitted. Each set of coefficients must begin on a new card and must be punched according to 26I3 format.

Format card

This F -type variable format card indicates the location of the raw scores (or ranks) on the data cards. This format may be punched in any of the columns on the card.

Data deck

These cards contain the data for each sample (or experimental condition) and must be punched in accordance with the format specified on the F -type variable format card (see above). For the Kruskal-Wallis test, the data are punched by sample with the data for each sample beginning on a new card. For the Friedman or Cochran test, the data are punched by subject (or group of matched subjects) with the data for each subject (or group of matched subjects) beginning on a new card.

Last card

If the user wishes to terminate the program, then the card immediately following the data deck must have the word FINISH

punched in columns 1 to 6. However, if the user wishes to analyze another set of data, then this card is a blank card, and the job deck is arranged sequentially (as described above) beginning with the problem card.

Output

The computer output for a given nonparametric test includes (a) the value of the corresponding statistic, i.e., H , χ^2 , or Q , (b) the number of degrees of freedom, and (c) the average rank (or condition mean) for each sample (or experimental condition). In addition, if the user chooses to have a post hoc trend analysis performed, and the null hypothesis is rejected by the appropriate nonparametric test, then the output includes the 95 per cent confidence intervals for the linear, quadratic, and cubic (if $K > 3$) comparisons.

Capabilities and Limitations

The program, which is written in FORTRAN IV, can handle a maximum of 30 samples (or experimental conditions) and 200 subjects per sample (or experimental condition). Jobs may be run sequentially as described above.

Availability

Copies of this paper and a source listing which includes input and output data for sample problems can be obtained by writing to Dr. James J. Roberge, Temple University, Department of Educational Psychology, Philadelphia, Pennsylvania 19122.

REFERENCES

- Cochran, W. G. The comparison of percentages in matched samples. *Biometrika*, 1950, 37, 256-266.
- Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 1937, 32, 675-701.
- Kruskal, W. H. and Wallis, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 1952, 47, 583-621.
- Marascuilo, L. A. and McSweeney, M. Nonparametric post hoc comparisons for trend. *Psychological Bulletin*, 1967, 67, 401-412.

A COMPUTER PROGRAM FOR TREND ANALYSIS IN A TWO- OR THREE-FACTOR EXPERIMENT WITH REPEATED MEASURES ON ONE OF THE FACTORS¹

JAMES J. ROBERGE
Temple University

THE use of experimental designs with repeated measures on one of the factors is extensive in the behavioral sciences. Moreover, the repeated measures factor in these designs, e.g., delay of reinforcement, length of isolation period, or dosage of a drug, often has levels which represent equally spaced intervals along an underlying continuum. Hence, additional information about the nature of the relationship between the repeated measures factor (treatment levels) and the dependent variable can be obtained by partitioning the treatment variation into nonoverlapping trend components, i.e., linear, quadratic, cubic, or quartic, through the use of orthogonal polynomials (Winer, 1962, pp. 353-369; Kirk, 1968, pp. 270-275).

The program discussed in this paper performs an analysis of variance for a two- or three-factor experiment with repeated measures on one of the factors. More importantly, it provides the researcher with the option of having a trend analysis performed on repeated measures factors of the type described above.

Input

The job deck set-up for each analysis is as follows:

Problem card

Columns 1-2 = number of levels of factor A.

3-4 = number of levels of factor B.

¹ The author gratefully acknowledges the support for this research which was provided by a Faculty Research grant funded by Temple University.

5-6 = number of levels of factor C.

7-9 = number of subjects per cell.

10 = trend analysis (1 = yes; 0 = no).

Coefficient cards

If the user opts to have a trend analysis performed on the repeated measures factor, then these cards contain the linear, quadratic, and cubic (if the number of levels of the repeated measures factor is greater than 3) coefficients of the orthogonal polynomials; otherwise, they are omitted. Each set of coefficients must begin on a new card and must be punched according to *1013* format.

Label cards

These cards (one per factor) contain the alphanumeric labels for factors A, B, and C (in a three-factor experiment), respectively. The label for each factor may be punched in any of the columns on the card.

Format card

This *F*-type variable format card indicates the location of the raw scores on the data cards. This format may be punched in any of the columns on the card.

Data deck

These cards contain the data for each subject and must be punched in accordance with the *F*-type variable format card (see above). Each card must contain the data for *one* subject. However, the use of more than one data card per subject is permitted.

Last card

If the user wishes to terminate the program, then the card immediately following the data deck must have the word *FINISH* punched in columns 1 to 6. However, if the user wishes to analyze another set of data, then this card is a blank card and the job deck is arranged sequentially (as described above) beginning with the problem card.

Output

The computer output includes (a) the labels for the factors, (b) an analysis of variance table, i.e., sources of variation, sums of

squares, degrees of freedom, mean squares, and F -ratios, which is presented in a form similar to that used by Winer (1962), and (c) matrices of means, standard deviations, and standard errors, for all main effects and interactions. Furthermore, if the user chooses to have a trend analysis performed on the repeated measures factor, then the analysis of variance table includes the trend components for the within subjects main effect, interaction(s), and error variation.

Capabilities and Limitations

The program, which is written in FORTRAN IV for processing by computers in the IBM 360 (or the CDC 6000) series, can handle a maximum of 10 levels for each of the factors and 10,000 observations. Well documented, it has variable names which are mnemonic, and correspond to the symbols used in Winer's computational formulas, to facilitate modification by users. Jobs may be run sequentially by introducing a new set of control cards and a new data deck as described above.

Availability

Copies of this paper and a source listing which includes input and output data for sample problems for two- and three-factor experiments can be obtained by writing to Dr. James J. Roberge, Temple University, Department of Educational Psychology, Philadelphia, Pennsylvania 19122.

REFERENCES

- Kirk, R. E. *Experimental design: Procedures for the behavioral sciences*. Belmont, California: Wadsworth, 1968.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

A STREAMLINED VERSION OF THE ALDOUS SIMULATION OF PERSONALITY¹

ROBERT A. LEWIS
Iowa State University

THE ALDOUS simulation of personality was developed by Loehlin (1962) at the University of Nebraska and was later refined at the University of Texas. The program was a weighted additive model of impression formation where values of stimuli, perceptual accuracy, and weighting factors could be manipulated (Loehlin, 1963). ALDOUS was basically a series of subroutines running in an environment (the main program) where inputs to the program represented stimuli from the environment. The model follows no formal theory of personality, but has been involved in several types of experiments (Loehlin, 1962, 1963, 1965, 1968).

The original program (Loehlin, 1965) consisted of approximately 750 assembly instructions for the Burroughs 205. Inputs and outputs for this version were entirely numbers. ALDOUS was later elaborated and rewritten in FORTRAN. The FORTRAN version required 624 source statements when made compatible with the IBM System/360 Model 65 at Iowa State University. When compiled in FORTRAN G, it occupied 23,810 bytes of storage. Inputs to this original FORTRAN version were numerical, and outputs were a series of sentences representing the model's reactions to the environment.

Streamlining ALDOUS

The original FORTRAN ALDOUS consisted of a main program (the environment), 10 subroutines, and four functions. Since all but three of these subroutines were called only once by any other routine,

¹ Computer time for this research was provided by a grant from the Dean of the College of Sciences and Humanities, Iowa State University, Ames, Iowa 50010.

the author determined that a considerable savings of computer time could be effected if most subroutines were inserted directly in the main program in the appropriate places. A complete description of each routine is available (Loehlin, 1963). The following routines were collapsed into the main program: STIMUL, CONSEQ. Routines RECOGN, EMOTN, ACTION, REACTN, LEARN and REPORT were condensed into the ALDOUS subroutine. These changes eliminated 10 sets of subroutine dimensioning statements without changing the logic of the program. More than 28 per cent of the statement numbers in ALDOUS were unused and were removed. There were also several cases where variables were packed into one array to pass from one routine to another and then unpacked upon arrival in the new routine. This practice was eliminated.

After this initial streamlining, the output volume was reduced. The original FORTRAN program used a series of numbers as input stimuli. These were retained. But the output consisted of a series of sentences constructed from concatenated phrases. In the modified program, verbal output was replaced by the numerical subscripts for the arrays which originally contained the phrases. This change reduced the output for each experimental trial from five lines of printed sentences to one line of digits which summarized the same information. It is now possible to replicate the design experiments in which hundreds of trials are run without generating an attendant mountain of output. Additionally, the numerical output can be directly analyzed statistically without further translation.

Testing the Revised Program

In replications of Loehlin's well-known experiment of development in two different environments (Loehlin, 1963), the modified FORTRAN ALDOUS and the original FORTRAN ALDOUS performed identically. The revised program, however, which is approximately 1,000 bytes smaller than the original, has an average run time of 31 per cent less. All experiments were conducted under a multi-programming environment.

This revision of the original FORTRAN ALDOUS retains the original input format, but uses a streamlined internal logic to produce a more compact output which summarizes a process identical to the original program. Copies of the flowchart and listings of the

revised program are available from: Robert A. Lewis, Department of Psychology, Iowa State University, Ames, Iowa 50010.

REFERENCES

- Loehlin, J. C. The personality of ALDOUS. *Discovery*, 1962, 23, 23-26.
- Loehlin, J. C. A computer program that simulates personality. In S. S. Tomkins & S. Messick (Eds.), *Computer Simulation of Personality*. New York: Wiley, 1963.
- Loehlin, J. C. Interpersonal experiments with a computer model of personality. *Journal of Personality and Social Psychology*, 1965, 2, 580-584.
- Loehlin, J. C. *Computer Models of Personality*. New York: Random House, 1968.

A COMPUTER PROGRAM FOR ESTIMATING THE POWER OF TESTS OF ASSUMPTIONS OF MARKOV CHAINS

ROBERT W. LISSITZ

University of Georgia¹

SILAS HALPERIN

Syracuse University

MARKOV processes are a popular model of longitudinal behavior (Gibbons, Halperin, and Lohnes, 1966; Lu, 1966; Atkinson, 1964; and Rapoport, 1969). They suggest themselves whenever a subject or sample of subjects is observed repeatedly across time and behavior is recorded as discrete categories. The Markov model is a very powerful one, allowing the user a great deal of parsimony as well as providing explicit predictions. The references given above have demonstrated the usefulness of this model for students of human behavior. Those readers looking for a well specified, dynamic, mathematical model should find their work of great interest. One of the most important advances in the study of individual differences has been the development of mathematical models and their appropriate and efficient utilization by the less mathematical but more empirical researcher. This model involves calculation of an initial vector of probabilities and a matrix of transition probabilities. Detailed descriptions of these parameters can be found in Kemeny and Snell (1960), and Anderson and Goodman (1957).

Certain testable assumptions must be made, though, before the above parameters can be calculated. These are assumptions regarding the order and stationarity of the stochastic process. Anderson and Goodman (1957) discussed these assumptions and presented

¹ The preparation of this manuscript was supported, in part, by P. H. S. traineeship MH-08258 from the NIMH, Public Health Service (1969-70). The senior author was in residence, during this time, at the Psychometric Laboratory of the University of North Carolina.

two test statistics and a body of sampling theory for evaluating their truth or falsity. One test is based on the likelihood ratio statistic and the other on a contingency statistic. In both cases these tests assume large sample size. Halperin (1966) studied the case where the null hypothesis is true and the test statistics limit (as sample size approaches infinity) to the chi-square distribution. The computer program described here concerns itself with the case where the null hypothesis is false and the experimenter is interested in the power of the statistical tests. In this case the test statistics limit (again, as a function of sample size) to the non-central chi-square distribution (Patnaik, 1949).

Very little is known, from a mathematical standpoint, about the power of these tests. Lissitz (1969) examined some characteristics of power using the methodology of the Monte-Carlo procedure. The computer program reported here grew out of this work.

This program allows a researcher to plan his study with regard to the power of these test statistics. He may use this program to select the number of states, the number of stages, and the sample size to obtain whatever level of power he desires. His problem would be simplified considerably if he could be certain of obtaining statistics which conform to the non-central chi-square distribution. For many research problems with small sample size this is not possible. Since no analytic solution has been derived for the sampling distribution of these statistics for small sample size, the researcher must depend upon Monte-Carlo procedures. This computer program performs the Monte-Carlo method and obtains an empirical solution. From this empirical sampling distribution the researcher is able to calculate the power of his tests and thus plan his research more carefully.

Input

The actual form of the card input is well specified by "comment" cards at the beginning of the program. Control of the program is by a series of cards prepared by the user. These cards specify the actual parameters of the population model (order I-stationary, order II-stationary, or order I-nonstationary) of interest and the critical values corresponding to the .10, .05, and .01 alpha levels.

Data Generation

The program uses the parameters of the hypothetical population to generate samples of data from which the test statistics can be

calculated. The following is the procedure for generating these random data vectors from an order one stationary Markov chain. Assume that there exists a chain whose initial probabilities are specified by the experimenter. It is desired to select a set of vectors from the totality of all possible vectors in this chain. Represent the m element (where m is the number of states in the model) initial vector of probabilities as a partition of a line interval of unit length. We can generate, by the power residue method, a random number between zero and one from a uniform distribution. This number must fall into one of the segments of the partitioned unit length interval. The segment into which this uniform random number falls will determine that person's starting state, or his position at time zero. In the limit, the proportion of random numbers which fall into segment i and are consequently classified into starting state i , will be equal to the probability of starting in state i as given by the i th element of the initial probability vector.

Once an initial element i of the data vector is randomly generated, we are in a position to use the i th row of the transition matrix to generate the next element of the data vector. Using the i th row in a manner identical to that of the initial vector we can generate a second element of the data vector which is consistent with the probabilities set down in that row. Again, in the limit, the proportion of people who start in state i and whose next state is generated to be j will be equal to the corresponding probability specified in the transition matrix. If we continue this procedure until the vector has $T + 1$ (where T is the number of transitions) elements, we will be able to interpret this datum vector as being randomly sampled from the specified Markov chain.

A similar procedure is used for generating the data vectors in the case of stationary-order two Markov processes and for order one nonstationary Markov processes. The only meaningful difference between the case of generating data vectors from the stationary order two population and the case discussed above is in the procedure used to initialize the process. In this situation there are two initial states to be determined before the transition matrix can begin prescribing probabilities. The program was written to input one initial vector and then to use it twice to generate two initial states. A random number is generated, and it is compared to the intervals of the initial probability vector. Thus, this process defines a state in

the manner outlined above. A second number is then generated, and it, also, is compared to this vector, and thus the second initial state is given. Together, these states prescribe the particular row and layer of the order two transition matrix.

The method of generating random data vectors for the nonstationary order one population is identical to that for the stationary order one situation except for the choice of the transition matrix. Instead of there being but one matrix there are T matrices and each of these is used in its turn. Again, the final result is a set of data vectors which conform, in the limit, to the population parameters.

Once a sample of data vectors is generated, the likelihood ratio statistic and the contingency statistic are calculated. This procedure is followed repeatedly, until an empirical sampling distribution is generated. The size of the sample and the number of samples are parameters specified by the program user.

In addition to the empirical values, a theoretical distribution is calculated, assuming that the limiting non-central chi-square distribution is appropriate. This theoretical distribution is a two parameter one (degrees of freedom and non-centrality value). The degrees of freedom are given by Anderson and Goodman (1957); and the noncentrality value, from the same article and from Lissitz (1969). The reader is referred to these sources for further discussion of this subject. Other methods for estimating the noncentrality value are possible but are not considered in this program.

Output

The output of this computer program is contained in two tables. One summarizes the distributions resulting from the Monte-Carlo sampling procedure and the theoretical noncentral chi-square subroutine. These are reported in the form of a grouped relative frequency table giving the proportions within each of the 14 intervals and the cumulative proportions. This is done, of course, for each of the two test statistics. Two summary statistics are provided: mean absolute difference between the theoretical and the empirical proportions, and the number of intervals with an empirical proportion outside the 95 per cent confidence limits of the theoretical proportion.

The second table in the output contains the power for the specific

Type I error rates specified on the last set of input cards. These error rates are presented for the two empirically determined distributions only. The first table allows (within the accuracy of interpolation) for calculation of power under the theoretical distribution, if this calculation is of interest.

Program Availability

The computer program is written in FORTRAN IV and currently operating on an IBM 360 model 50. The program calls for two subroutines from the Scientific Subroutines package (IBM, 1968). A listing of the program is available upon request from R. W. Lissitz.

REFERENCES

- Anderson, T. W. and Goodman, L. A. Statistical inference about Markov chains. *Annals Mathematical Statistics*, 1957, 28, 89-110.
- Atkinson, R. C. *Studies in Mathematical Psychology*. Stanford, California: Stanford University Press, 1964.
- Gribbons, W. D., Halperin, S., and Lohnes, P. R. Applications of stochastic models in research on career development. *Journal of Counseling Psychology*, 1966, 13, 403-408.
- Halperin, Silas. Markov models for human development: statistical testing and estimation. Unpublished dissertation, University of Buffalo, 1966.
- IBM Corporation, Scientific Subroutines. Technical Publications Department, H20-0205-3, 1968, White Plains, New York.
- Kemeny, J. G. and Snell, J. L. *Finite Markov Chains*. Princeton, New Jersey: D. Van Nostrand Company, Inc., 1960.
- Lissitz, Robert W. Testing assumptions of Markov chains: Empirical and theoretical distributions under the alternative hypothesis. Unpublished dissertation, Syracuse University, 1969.
- Lu, K. H. A path-probability approach to irreversible Markov chains with an application in studying the dental caries process. *Biometrics*, 1966, 22, 791-809.
- Patnaik, P. B. The non-central χ^2 and F distributions and their applications. *Biometrika*, 1949, 36, 202-232.
- Rapoport, A., Effects of observation cost on sequential search behavior. *Perception and Psychophysics*, 1969, 6, 234-240.

SUBROUTINE TO DECODE IBM 1230 DATA

JOHN W. MENNE

Iowa State University

JOHN E. KLINGENSMITH

Arizona State University

THIS FORTRAN IV/360 decoding procedure uses the machine bit form of data read in alphanumeric format to calculate in binary arithmetic an index which is used to address the stored vector of the decoded values. Decoding in this way seems to improve over routines such as described by Veldman (1967, pp. 167-169) by about a factor of five. The subroutine is limited to 16/32 bit machines, but the procedure is adaptable to other machines. Other features are:

1. Dimensions are supplied by the calling program, as in IBM supplied subroutines.
2. The same storage vector is used for the alphanumeric input and the decoded (integer*2 or integer*4) output. Of course, on input this vector is only partially filled with coded data.
3. Valid characters which are not part of the IBM 1230 code will not be returned to the calling program but will be printed with identification.
4. The routine requires 922 bytes.

Source deck and listing, including usual calling statements, are available from the Student Counseling Service at Iowa State University.

REFERENCE

- Veldman, Donald J. *FORTRAN programming for the behavioral sciences*, New York: Holt, Rinehart and Winston, 1967.

COMPUTER PROGRAMS FOR RANK ANALYSIS OF COVARIANCE

EDWARD P. LABINOWICH

San Fernando Valley State College

JAMES K. BREWER

Florida State University

IN many experimental studies the nature of the sample dictates the adoption of nonparametric statistical techniques, since these are based on less stringent assumptions and are often more generalizable than parametric techniques. FORTRAN IV programs are provided in this paper for the non-parametric one-way analysis of covariance based on the procedure proposed by Quade (1967).

In Quade's original paper, various methods were discussed for the comparison of two or more samples with respect to a response variable Y in the presence of a concomitant variable (covariate) X —a situation for which the usual analysis method is a standard one-way analysis of covariance. Two distinct methods have been developed by Quade—one for the analysis of covariance by ranks for one covariate and another for the case of two covariates. A separate FORTRAN IV program has been written for each technique. Despite a distinct contrast in the specific details for the determination of the variance ratio by each method, the input and output for the computer programs are basically the same.

Input

Each job deck contains control cards which record the total sample size and the size of each group. The data decks are punched according to a format card and each data card provides the following information for each subject: the Y score, the Y rank, the X score and the X rank (X_1 and X_2 when there are two covariates).

Output

The computer output includes for each subject: The Y ranks and the X ranks corrected for the median, the predicted Y rank, and the residual from the regression of Y ranks. Values for the variables at each stage of computing the final variance ratio, e.g., the sum of squares, are included in the output.

Capabilities and Limitations

The programs are documented and employ variable names which correspond mnemonically to the symbols assigned by Quade in his computational formulas. Both programs are applicable only to two sample problems as illustrated in Quade's examples.

Availability

A print-out and sample output of each program can be obtained by writing Dr. E. Labinowich at the School of Education, San Fernando Valley State College, Northridge, California 91324, or Dr. James K. Brewer at the Department of Educational Research and Testing, College of Education, Florida State University, Tallahassee, Florida 32306.

REFERENCE

- Quade, Dana. Rank analysis of covariance. *Journal of the American Statistical Association*, 1967, 62, 1182-1200.

BOOK REVIEWS

MAX D. ENGELHART, Editor

Duke University

HENRY MOUGHAMIAN, Assistant Editor

City Colleges of Chicago

<i>Brown's Principles of Educational and Psychological Testing.</i>	299
ROSS E. TRAUB AND C. W. FISHER	
<i>Cohen's Statistical Power Analysis for the Behavioral Sciences.</i>	303
ROSS E. TRAUB	
<i>Cramer, Herr, Morris, and Frantz's Research and the School</i>	307
<i>Counselor. R. B. SIMONO</i>	
<i>Hays and Winkler's Statistics: Probability, Inference, and</i>	310
<i>Decision. JAMES A. WALSH</i>	
<i>Heermann and Braskamp's Readings in Statistics for the</i>	312
<i>Behavioral Sciences. LEWIS R. AIKEN, JR.</i>	
<i>Kleinmuntz's Clinical Information Processing by Computer:</i>	314
<i>An Essay and Selected Readings. RICHARD WOLF</i>	
<i>Lemaine and Lemaine's Psychologie Sociale et Experimenta-</i>	316
<i>tion (Experimental Procedures in Social Psychology).</i>	
ROBERT SMITH	
<i>La Recherche en Enseignement Programme—Tendances Ac-</i>	317
<i>tuelles (Programmed Learning Research—Major Trends).</i>	
ROBERT SMITH	

Frederick G. Brown. *Principles of Educational and Psychological Testing*. Hinsdale, Ill.: The Dryden Press, 1970. Pp. vii + 468. \$9.95.

The content of Brown's book on educational and psychological testing is summarized fairly well by the chapter headings: In order of appearance are chapters discussing the nature of measurement, test development, reliability, validity, scores and norms, how to combine scores, measurement in the domains of achievement, aptitude and personality, and, finally, problems and trends in testing. A brief account of descriptive statistics is given in an appendix to the first chapter. The book seems best suited to training test users, not test developers. This judgment is supported by the fact that relatively little space (approximately 20 pages) is devoted to item writing and item analysis whereas a relatively large amount of space (approximately 165 pages) is used to describe different types of standardized instruments and how they may be evaluated. As advertised in the preface, Brown has successfully avoided the temptation to include tedious catalogues of standardized instruments. Instead, different types of standardized measures are illustrated through reference to familiar and widely-used tests and questionnaires. An excellent guide for evaluating tests is presented in outline form in Chapter 9. It will be of undoubted assistance to those readers who have the task of selecting a standardized instrument for a particular purpose.

The book has several other attractive features. It is well written in an informal style that should appeal to students. It includes important topics not ordinarily found in introductory texts. Two examples of such topics are (1) assessing the accuracy of selection decisions that are made on the basis of how examinees score on a predictor test, and (2) discriminant analysis. These, as well as more familiar topics such as factor analysis and multiple-regression, are developed with a view to having the reader understand the purpose and rationale of the technique. Numerical examples, standing separately from the text, are provided to illustrate each statistical technique. But mathematical derivations are omitted to meet Brown's stated assumption that not all readers will have done courses in statistics. Also worthy of note is the study guide for the book. It contains previews of each chapter and questions that will help focus the student's attention as he reads.

The book's positive points have been given first emphasis in order to create a suitable perspective from which to view the negative

comments that follow. These are offered to alert potential users of the book to difficulties that can be surmounted through class lectures and discussions. Very occasionally Brown lapses into the use of questionable terminology and imprecise language. Reliability is defined parenthetically as "degree of inconsistency" (pp. 52-3) but elsewhere as degree of consistency. Brown interprets standard error of measurement (p. 85) and standard error of estimate (p. 115) as standard deviations of a normal distribution without explicitly noting the assumption underlying this interpretation. And on two occasions (p. 113 and p. 192) the predicted score that results from the application of a linear regression equation in one predictor is described as the mean of the criterion scores made by persons with the same predictor score. Brown should also have stated that this description provides only a convenient approximation to reality.

In at least one case Brown can be faulted for failing to develop a concept fully. The notion of optimum test discrimination is treated as follows: "For tests designated to discriminate between students, a mean score slightly higher than 50 per cent of the maximum possible score is optimal (with an approximately normal distribution)" (p. 274). This assertion may represent Professor Brown's experience with what is possible in practice. Ideally, a rectangular distribution of scores would be desired if the situation required optimum discrimination among all the students tested. Rectangular distributions are, of course, rarely, if ever, observed. But tests can probably be constructed to yield distributions of scores that are more platykurtic than the normal distribution. Such tests would provide relatively better overall discrimination among students than the tests of which Brown speaks.

Perhaps more serious are a contradiction and a bit of misinformation that appear in the book. The contradiction occurs with respect to the notions of interval measurement and standard score scales. Brown gives the impression (p. 9, p. 163 and pp. 169-170) that by transforming raw scores to standard scores, measurement on the level of an interval scale may be achieved. He does so after having raised the question (pp. 7-8) of whether interval measurement is possible for most educational and psychological variables. In addition, he fails to acknowledge that if standard scores are on an interval scale then so are the corresponding raw scores inasmuch as they are merely linear transformations of the standard scores.

The bit of misinformation is conveyed in the assertion (p. 79) that factor analysis can be used to determine whether a test is unifactorial. Brown glosses over two problem factor analysts face: (1) deciding what type of correlation coefficient to compute when test items are dichotomously scored, and (2) determining the rank of the matrix being factored. Brown fails to acknowledge that if these problems are solved and it is found that a matrix has a rank of one,

then a necessary but not sufficient condition of unidimensionality has been met (Lord and Novick, 1968, p. 382).

An attempt could be made to excuse many, if not all, of the foregoing complaints with the explanation that to overcome them would require the use of more sophisticated mathematics and more involved verbal explanations than is warranted in an introductory text. It is clear that we would argue this point. But potential users will have to make up their own minds about how intellectually honest an introductory text can be and still be written understandably.

On two other occasions in the book, Brown made what we feel are arguable decisions. One involves the treatment of reliability. The Kuder-Richardson coefficients are presented as indices of homogeneity, not of reliability. Homogeneity is a concept with a long and confused history. Brown tries to avoid most of the problems associated with it by ignoring the work of Louis Guttman, among others, and by defining homogeneity as ". . . consistency of performance over all items on a test" (p. 77). But this definition fails to clarify the concept for the reader. In particular, we were left wondering whether homogeneity by this definition could mean that all the items of a test have linearly related true scores of the special type referred to by Lord and Novick as "essentially tau-equivalent" (Lord and Novick, 1968, p. 50). If it does, then the KR-20 estimate of reliability for a set of items that satisfy the definition equals the proportion of true score variance in the variance of observed scores, and this is how Brown defines reliability. This relationship between reliability and one conception of homogeneity provides an exception to the assertion ". . . that reliability will always be greater than homogeneity" (p. 81). Another, possibly unfortunate, aspect of this approach to the treatment of reliability is that it leads Brown to interpret split-half coefficients as estimates of equivalence when in fact they may also be regarded as estimates of internal consistency. This follows from the well-known fact that the mean of all possible split-half coefficients for a test, each computed according to what Brown refers to as Guttman's formula (p. 67), is equal to the value of KR-20 for the test.

The other decision about which Brown and we disagree is the use of the correction-for-guessing formula. He concludes with respect to the use of the formula ". . . that the burden of proof would fall on the proponents of correcting for guessing" (p. 273). We agree that this would be the case where the instructions to the student do not specify what he is to do when he encounters a question that cannot be answered with confidence. But this is bad practice. The instructions should tell the student either to guess or not, and in the latter case, it is surely necessary to motivate him not to guess by informing him of a penalty for wrong answers or a reward for omitted questions. When reward or penalty instructions

and the corresponding correction-for-guessing formula are used, the resulting test scores can be expected to be more reliable and more valid than if the student is encouraged to guess as Brown recommends.

On the basis of the strengths and weaknesses we identified in the book, we conclude that Brown has, in fact, produced a book that is very good in many respects. It stands as a testimonial to Brown's breadth of knowledge and eclecticism. Many instructors will undoubtedly find *Principles of Educational and Psychological Testing* useful for their courses.

Having offered this conclusion, it is possible for us to deal with the broader issue of the contribution the book makes to the field of educational and psychological measurement. Here we also comment on the state of the field as it now exists. Brown's book is one of many to appear in the last few years, each providing a restatement of the discipline in fairly traditional terms. This leads us to ask whether these books were all necessary? Are "new" books being written more to make singular contributions to the field or to satisfy the desire of publishers for fairly traditional, eclectic, and therefore salable products?

The *American Scientist* recently carried an article by Paul A. Weiss (1970) entitled *Whither life science?* One of the observations Weiss made in his article concerns the textbooks that existed in biology when he started working in that field some fifty years ago. Then, Weiss says, "Textbooks were few, comprehensive, original, unique, almost everyone of them bearing the signature of a master; but obviously there were wide gaps between the areas they covered. Yet they had one important feature in common: they tried, some more than others, to balance overindulgence in their particular speciality by pointing up the place and context of that speciality within the continuum of the living world. In this way we became aware of both the fundamental interconnectedness of all aspects of life and the appalling dearth of concrete knowledge about the interconnections, provisionally labelled by symbolic terms" (p. 158). Although Weiss does not say it, we sensed that he believes most present-day textbooks do not present biological knowledge in the unique and original way of older textbooks. More important is his implication that modern textbooks fail to define the problems or describe the goals of biology.

It can be argued, we think, that much the same situation exists in the field of educational and psychological measurement. Among recent new textbooks, one looks in vain for highly original contributions. Many new textbooks, Brown's included, are very comprehensive in their coverage of measurement topics, but most teachers of educational and psychological measurement will know of well established books that are at least as comprehensive. And no recent

book we have seen would really satisfy Weiss in that it adequately states the problems and identifies the goals of educational and psychological measurement as it is practiced today. Perhaps what is needed to advance the field is for fewer people to spend their time writing textbooks and for more to devote their time to research and to the development of new measurement concepts and techniques. Then, in time, truly "new" textbooks could be written.

REFERENCES

- Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.
Weiss, P. A. Whither life science? *American Scientist*, 1970, 58, 156-163.

ROSS E. TRAUB AND C. W. FISHER
The Ontario Institute for Studies in Education

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press, 1969. Pp. xv + 415. \$13.50.

The power of a statistical test, according to the classical theory of hypothesis testing advanced by Neyman and Pearson, is the probability of rejecting a null hypothesis in favor of an alternative hypothesis when the alternative one is true. Three factors affect power: the probability of rejecting a null hypothesis when it is true (hereafter symbolized by α), sample size, and the standardized magnitude of the difference between the parameter value specified in a null hypothesis and its true value (hereafter referred to as "effect size"). Power increases as α or sample size or effect size increases. The advantage to a scientist who knows all this is that, in theory at any rate, he can plan his experiments so as to achieve a satisfactory degree of power. Now, the main factor amenable to manipulation in planning experiments is sample size. Cohen's book is designed primarily to help a researcher with foresight determine the number of subjects he needs to achieve a desired degree of power.

Cohen begins the book with a chapter explaining the concept of statistical power and discussing several of the problems that had to be solved in the process of preparing the book. More about these later. The next seven chapters, comprising over 90 per cent of the book, contain descriptions of statistical tests and tables of results. A separate chapter is devoted to each of the following topics: the t test for means, the test for the significance of a product-moment correlation coefficient, the test for the significance of a difference between product-moment correlation coefficients, the test that a proportion is different from 0.5, the test of the significance of a difference between two proportions, the chi-square test, and

the F test as employed in fixed-model analyses of variance and covariance. A final chapter describes the computational procedures used in generating the results.

Approximately 180 pages of the book are devoted to power and sample size tables. (In addition, there are about five pages of miscellaneous tables.) For specified levels of α , the power tables contain estimates of power for combinations of selected sample sizes and selected effect sizes. Each table also indicates how large an observed effect would have to be for it to be statistically significant, given a specified α -level and sample size. Consequently, the power tables can be used as an aid in significance testing. The sample-size tables, on the other hand, indicate how large the sample must be, again for a particular α -level, in order that a specified degree of power will be achieved with respect to a specified effect size.

In preparing the tables, Cohen considered three α -levels: 0.01, 0.05, 0.10. Consequently, three power and sample size tables are reported for each variation of each statistical test that was studied. Two factors determined which variations would be included in the analysis: (1) For some tests, nondirectional (two-tailed), as well as directional (one-tailed), alternative hypotheses could be of interest. In such cases, separate sets of three power or sample-size tables are provided for each type of alternative hypothesis. This means that the amount of information available about such tests is increased because, to a reasonably close approximation, the tabled power (sample-size) values for nondirectional alternative hypotheses may be regarded as power (sample-size) values of directional alternative hypotheses at one-half the reported level of α . On the other hand, the tabled power (sample-size) values for directional alternative hypotheses may be regarded as power (sample-size) values of nondirectional tests at double the reported level of α . (2) For the chi-square test, sample size is independent of the number of degrees of freedom for the test. A similar independence exists in the F test of fixed-model analysis of variance and covariance; the independence is between the number of degrees of freedom for the lesser (expected) mean square (which is dependent on sample size) and the number of degrees of freedom for the greater (expected) mean square (which is *not* dependent on sample size). In both these cases separate sets of power and sample-size tables are provided for each different number of (sample-size-independent) degrees of freedom that is considered.

As indicated previously, Cohen had to solve several problems in performing the reported power analyses. One problem was to define a metric-free index of effect size for each statistical test, an index that would not reflect the type of data accumulated in a particular study. Where possible, Cohen chose effect-size indices that were related through the concept of "... proportion of variance accounted

for in the dependent variable" (p. 12). This enabled him to solve another problem, that of providing a rough basis for judging the magnitude of an effect size. A small effect is defined as one that accounts for approximately one per cent of the variance in the dependent variable, a medium-sized effect for nine per cent, and a large effect for 25 per cent. Another problem was to illustrate how the results of the power analyses can be used. Cohen handled this problem by considering different "cases" involving a particular test and providing one or more illustrative examples of each case. For example, in the chapter on the t test for means, five cases are considered: the difference between the means of two samples drawn from populations with equal variances when the sample sizes are equal (case 0) or unequal (case 1), the difference between the means of two samples drawn from populations with unequal variances when the sample sizes are equal (case 2), the difference between the mean of one sample and an hypothesized value of the mean (case 3), and the difference between two means when the observations yielding the first mean are correlated with the observations yielding the second (case 4). The power and sample size tables for the chapter on the t test were constructed using the assumptions of case 0. Therefore, the tables do not necessarily apply to cases 1-4. Cohen indicates when they may reasonably be applied to cases 1-4, shows how to use them in those cases, and provides an idea of how large the discrepancies may be between the tabled results and the true results.

Statistical power analysis for the behavioral sciences could appeal to a wide range of behavioral scientists. It may be read by anyone who has mastered the statistics found in introductory textbooks on the subject. Cohen does not trouble the reader with detailed mathematical derivations. The writing contains enough intentional redundancy to enable the reader to test his understanding of concepts as he progresses. And the many examples contained in the book promote understanding and help the reader to use the results.

Despite these positive characteristics, it is doubtful that the book will be widely used. The reason for such a pessimistic prediction is that power analysis is usually very difficult to do in the planning stage of experiments because investigators do not often have a good idea of the effect size that may be expected. When effect size is unknown, it is impossible to determine the sample size required to achieve a given power level. Of course the problem may be attacked in a different and possibly more useful way (Glass and Stanley, 1970, pp. 287-288). In this approach, the investigator first determines the size of the largest sample he can afford to take. Then the problem is to determine whether this sample size is large enough to provide a satisfactory degree of power in the event that the effect size is as small as it could possibly be and still be interest-

ing. Such an approach may reveal that a smaller sized sample would maintain a satisfactory power level or it may indicate that the investigator's resources are so small that he cannot reasonably expect to detect an effect even if it is reasonably large. This type of thinking can be assisted by the information contained in *Statistical Power Analysis*.

Another reason why the book may not be used extensively is that it is not as comprehensive as it might have been. Many behavioral scientists employ nonparametric statistics in their research. For them, the utility of the book would have been enhanced had analyses been made of tests involving rank-order correlation coefficients or the Mann-Whitney U test, to cite two examples.

There are other aspects to the book that may annoy the reader. One thing that disturbed me is that Cohen failed to support many of his statements with appropriate references to the literature. For example, in his discussion of the test of the significance of a product moment correlation coefficient, Cohen makes the following assertion: "However, when significance tests [of the product moment correlation coefficient] come to be employed, assumptions of normality and homoscedasticity are formally invoked. Despite this, it should be noted that, as in the case of the t test with means, moderate assumption failure here, particularly with large n , will not seriously affect the validity of significance tests, nor of the power associated with them" (p. 72). No reference is given in support of the assertion. Why is the scholarship of writers of statistics books not held accountable in the same way that the scholarship of other scientists is? Another, relatively small point, that may disturb the statistical sophisticate, is Cohen's use of bold face Latin letters to symbolize both population parameters and sample statistics. The symbol for the sample statistic differs only in having the letter s as a subscript.

Finally, mention should be made of the fact that *Statistical Power Analysis* contains its share of errors and ambiguities. The range of these faults, both in terms of type and degree of seriousness, has been documented by McNemar (1970). My favorite is the statement that a nondirectional test has less power than a directional test "... provided that the sample result is in the direction predicted. Since directional tests cannot, by definition, lead to rejecting the null hypothesis in the direction opposite to that predicted, these tests have almost no power to detect such effects" (p. 5; essentially repeated on p. 39).

Educational and psychological researchers who attempt to use the logic of power analysis in planning their experiments and who employ the tests discussed in *Statistical Power Analysis* will want to add the book to their reference collections. They are particularly encouraged to do so if they find tables easier to use than nomo-

graphs. Beyond that, *Statistical Power Analysis* may have some utility as a supplementary reference for advanced courses in educational and psychological statistics courses. But teachers are well advised to read both the book and McNemar's review with care before making class assignments.

REFERENCES

- Glass, G. V and Stanley, J. C. *Statistical methods in education and psychology*. Englewood Cliffs, N. J.: Prentice-Hall, 1970.
McNemar, Q. Potential power for your static stat. *Contemporary Psychology*, 1970, 15, 452-453.

ROSS E. TRAUB

The Ontario Institute for Studies in Education

Stanley Cramer, Edwin L. Herr, Charles N. Morris, and Thomas T. Frantz. *Research and the School Counselor*. Boston: Houghton Mifflin Company, 1970. Pp. vii + 202. \$3.50 (paperback).

In the introduction to this text, C. Gilbert Wrenn commented that this book was "for first-year graduate students or counselors on the job who have little research knowledge" and that it was written in a clear and straightforward style. These are accurate descriptions for the work and this writer would add that nothing has been lost by presenting the material in such a readable fashion. Throughout the text the authors attempt to convince school counselors that research is a necessary part of their responsibility and in so doing, the authors help to remove much of the aura surrounding the term "research".

The first few chapters present some starting points that the counselor might utilize in launching certain research activities in the particular setting in which he finds himself. Adequate descriptions were given of the types of research that the counselor might engage in, such as descriptive research, environmental assessment, follow-up studies (what happens to students after they leave school), and finally research through which the counselor comes to some conclusions about the effectiveness of particular counseling programs and techniques. However, the section on the "asking" of questions opened a very crucial area but could have been expanded upon much more. The chapter on elementary descriptive techniques was effective and particularly useful since the authors attempted to combine it with some comments about the use of data generally found in a school setting (e.g. cumulative records). Topics covered were frequency distributions, measures of central tendency, standard deviation, normal curve, and standard scores. The discussion of standard scores was somewhat compact and might be difficult for someone who is first being introduced to such a concept, but this is

quite a minor criticism of this particular chapter. Also practical and clear was the discussion on the construction and use of norms noting that the counselor should be encouraged to construct those which are most appropriate for the sample with which he is working.

Chapters four and five led the reader into a consideration of the area of expectancy tables and correlation techniques. The reader was presented with the rationale for expectancy tables and comments were made about their use in communicating to students, parents, and teachers the predictive quality of some of the information available to students. This writer was pleased to see that the authors urged counselors to see that the interpretation made to students on the basis of an expectancy table "is not a substitute" for counseling but just another source of information to be used in the process and the authors supported this contention by providing some discussion of the cautions to be employed when using such tables. The chapter on correlation techniques might be a leap into the unknown for some readers who have no prior experience with this concept. Also this chapter on correlation might have been tied in better with the previous discussion on expectancy tables. But after this start the chapter slows down enough so that even the counselor inexperienced in statistical techniques should be able to develop a sound understanding. Not too practical was the presentation on the various types of correlation techniques determined by the variable or variables being researched. It appeared doubtful in this writer's opinion that most counselors would venture this far into the data. But the writers did cover well the implications for the utilization of correlation techniques. Reliability and validity seemed to get very little attention in this section and if the reader has not thoroughly examined these applications of correlation to measurement he will be in for a very difficult time in this section.

The authors set aside one chapter to introduce the reader to the new and developing research area of environmental assessment. Because the field is just beginning to get off the ground it is often difficult to find much written on research rationale and methodology in this area and therefore, this chapter should be a welcome sight to school counselors. It is obvious that every day it is becoming more important to have some evidence as to the impact of the school milieu upon the individual student. Described were some instruments available for assessing the environment of such settings as high schools, colleges and universities and most important seemed to be the discussion on the means of taking on such a research task. The counselor probably has not always been able to meet the expectations of institutional researcher for many reasons but this chapter should give impetus to those so inclined.

Chapters seven and eight dealt with traditional research areas such as the overall evaluational guidance programs and followup

studies. Probably the most important research area in counseling will continue to be the evaluation of counseling services. Sensitive questions were raised early in the chapter but appeared to be the kind of issues counselors will have to respond to in the course of their research activities. "Does counseling do what it purports to do?" essentially was the basic question. The criterion program in such research was clearly discussed and examples of criteria were given which might have practical significance in school setting. Also mentioned were realistic problems of sampling and controls. Overall this chapter was quite thorough and reflected the importance which the authors appropriately assigned to this counselor responsibility. In a following chapter the authors also dealt with the formalized followup of school dropouts and graduates. On the strength of some very basic followup research in guidance (e.g. J. W. M. Rothney's work) the authors gave the school counselor a good base upon which to make decisions concerning followup. This is a very practical chapter and covered very well the alternatives of questionnaires and interviews in such activities. Important also was not only a discussion of the procedures for planning and implementing a program of followup research but the authors wisely included some cautions to be observed while interpreting the data.

Chapters nine and 10 dealt with class rank, academic average and the measurement of aptitude and opinion. Procedures were shown for using results of class ranking and academic averaging in order to communicate valid descriptions of designated groups in a school setting. The discussion of measurement of opinions and attitudes was not unique and is the kind of discussion often available in most basic measurement texts.

Certainly one of the most useful chapters was the one on studies of school dropouts and the importance of such data to school systems when properly used. Many practical suggestions were given and discussed thoroughly were the four general types of research on school dropouts: the reasons approach where brief responses are gathered, the case study approach, the factor approach which included the examination of such things as aptitude, socioeconomic status, etc., and finally a broad social systems approach.

It was mentioned earlier by this writer that environmental assessment is a new and recent development in school systems but also relatively new on the scene is the use of data processing at many levels of education. The chapter on the counselor and data processing focused on the need that the counselor will have to utilize such procedures more and more. If some schools are not presently involved in extensive data processing, the time is almost upon us when this will be a common aspect of record keeping at all levels of educational operation. The authors covered well both pencil-and-paper methods as well as automatic data processing. This discussion

was quite detailed and systematic but this reviewer found it to be very interesting and highly readable. Again the authors reflected upon all the important implications such as uses in the storage and retrieval of occupational and educational information utilized in all stages of individual development.

What better way to conclude a work on practical research for the counselor than to include a look at the research activities of such men as Donald Super and John Rothney. Included was Super's *Career Pattern Study*, Rothney's *Guidance of American Youth*, and *Guidance Practices and Results* and Krumboltz's *Revolution in Counseling*. However, Krumboltz is at an early stage in his research career and the counselor will probably have to wait some years to realize the practical significance of his work in comparison to that of Rothney and Super.

When this writer first opened this text he felt that this might be just another organization of bits of basic statistics texts and measurement texts but this was not true for this particular book. Both the school counselor and graduate students beginning their programs in counselor education should welcome this work both as a tool for learning and also as an excellent source of information to be kept close at hand in one's work setting.

R. B. SIMONO

University of North Carolina at Charlotte

William L. Hays and Robert L. Winkler. *Statistics: Probability, Inference, and Decision*. New York: Holt, Rinehart and Winston, 1970. Volume I, pp. xviii + 650, \$10.95. Volume II, pp. xiv + 320, \$8.95.

It has become more and more difficult to come up with an introductory textbook in applied statistics that has a truly new look, but Hays and Winkler have done just that with their two-volume *Statistics: Probability, Inference, and Decision*. The novelty of their approach lies in a unique combination of four elements: (a) a comprehensive treatment of the fundamental ideas of probability and probability distributions; (b) the direct use of these ideas to develop insight into the theory of classical statistical inference; (c) the thorough integration of Bayesian thinking and decision theoretic concepts into the inferential process; (d) a concluding section (the second volume) on statistical methods. The resulting mixture is an exciting one that gives an excellent elementary introduction to decision theory, but the work as a whole gives one the feeling of a lack of satisfactory closure. Basically, the two volumes do not seem to form a set. The material in Volume I on probability and classical inference is clearly necessary for both the chapters in Volume I on Bayesian inference and

decision theory and for the statistical methods described in Volume II. However, Volume I builds up to an understanding of decision theory as a sort of ultimate inferential process, but then does not effectively carry this outlook over to the statistical methods in Volume II. It seems likely that the student who uses both volumes for a year's course in applied statistics for behavioral scientists will come out with a schizoid feeling about decision theory and statistical methods rather than with an appreciation of the need for the development of methodology to routinely generalize decision theoretic thinking to the use of statistical methods.

Each of the two volumes has distinct strengths and weaknesses that deserve comment. The biggest fault of Volume I is that it attempts to deal with continuous distributions as well as discrete ones while assuming both a background in calculus and no background in calculus (through the use of "heuristic" explanations). The result is sometimes like using the *Encyclopaedia Britannica* as a first grade reader: the student may understand what is being talked about if the teacher supplies the vocabulary word by word, but he cannot do anything with the material by himself. This criticism holds principally for the mathematical sections and subsections. Hays' talent for readable, detailed exposition is everywhere evident and the major concepts are generally very clear and understandable. Appropriately, then, the greatest strength of Volume I is the clarity with which the authors show how decision theory provides tools for combining prior information about some phenomenon of interest with sample information to reach a decision that is best in some particular way, such as maximizing gain, or minimizing loss, or some more complex criterion.

Volume II appears to be potentially less useful than Volume I in that it is not really compatible with Volume I and yet does not stand alone. The primary difficulty is that the selection of statistical methods covered is too narrow although often quite deep. The very complete development given regression and correlation and the much-better-than-average section on sampling are laudable, but the discussion of the principles of experimental design is both too brief and too scattered among the computational details of various experimental designs. Moreover, notions of multiple comparisons are almost totally lacking. The treatment given nonparametric methods is admirable, however. This latter section is well-organized and quite modern in outlook.

In summary, although Hays and Winkler have produced a new approach to elementary applied statistics, it is questionable whether the effort as a whole has much to commend it. The volumes are comprehensive—even massive—and well written. They cover in excellent fashion some materials on decision theory not heretofore available in elementary form. But, on the debit side of the ledger,

breadth and depth of coverage of statistical methods are not well-balanced and the authors consistently beg the question of teaching a mathematically-based discipline without requiring a sound training in mathematics. As other authors who have faced this question have done, Hays and Winkler talk *about* the use of mathematical procedures extensively without ever requiring the student to take the plunge and *do* some mathematics. This criticism is not Utopian in tenor: a great many interesting problems in statistics *per se* can be tackled with only the elementary calculus as background. Nevertheless, Hays and Winkler have done an excellent job in Volume I of presenting the basic notions of decision theory and Bayesian reasoning. In Volume II, the sections on correlation and regression and on nonparametric methods are exemplary. For these reasons, *Statistics: Probability, Inference, and Decision* should enjoy some deserved popularity during the period in which psychology makes up its mind that mathematics is the language of science and joins other disciplines in demanding of its students a thorough grounding in its essential branches.

JAMES A. WALSH
Iowa State University

Emil F. Heermann and Larry A. Braskamp (Eds.). *Readings in Statistics for the Behavioral Sciences*. Englewood Cliffs, N. J.: Prentice-Hall, 1970. Pp. ix + 419. \$4.95.

This collection of thirty-one papers, eleven of which were published originally in the *Psychological Bulletin* and the majority of the remainder in recent issues of psychological and educational journals, is designed as a supplementary text for undergraduate and graduate courses in statistics or as a textbook for statistics seminars. Since students in the required basic and intermediate statistics courses have little enough time to study a core text and work the assigned exercises, this book will probably find a more receptive audience in graduate seminars. Although the readings and associated examples are aimed primarily at professionals in education and psychology (so perhaps the term "behavioral" in the title is a bit pretentious), most of the issues and problems discussed are of a general methodological nature.

Predictably for a first edition, the book contains some typographical errors and oversights, but usually these do not detract. One exception is on the back flyleaf, where someone neglected to include a description of the contents of parts I and II. Also there is a surplus of papers on certain issues, for example the question of the relative merits of parametric and nonparametric techniques. And a glance at a list of statistical reviews and notes appearing in the *Psychological Bulletin* over the past 25 years reveals that many

important papers were omitted. In addition, certain older papers and a more thorough discussion of historical controversies would have been of interest to the reviewer. However, the editors had to make choices, and by and large they made excellent ones.

The book is divided into six sections: I. History of the Application of Statistical Methods (2 papers), II. Parametric vs. Non-parametric Statistics (6 papers), III. Randomization (4 papers), IV. Testing Statistical Hypotheses (6 papers), V. Special Topics in Analysis of Variance (9 papers), and VI. Correlation and Regression (4 papers). An introductory overview of the papers is given at the beginning of each section.

The two papers in the first section, the first by Helen Walker and the second by Jerzy Neyman, deal with the contributions of Karl Pearson and R. A. Fisher and highlight the long-term argument between these two giants as well as the contrast between the Neyman-Pearson and Fisherian concepts of hypothesis testing and interval estimation. The majority of the papers in the second section, which is concerned with the issue of parametric vs. nonparametric statistics, are relatively short. A provocative defense of the controversial matter of levels of measurement (Stevens, 1951) is presented in S. S. Stevens' paper on "Measurement, Statistics, and the Schemapiric View." Finally, Donaldson's discussion of the robustness of the F test is perhaps the most technical paper in this section.

The main issue considered in the third section ("Randomization") is the importance of the randomization assumption underlying tests of hypotheses and the extent to which the sampling distributions associated with these tests approximate those of the exact randomization tests. The empirical results presented in the Baker and Collier papers on the effects of skewness, kurtosis, and number of observations on the F ratios in the completely randomized design and the effects of block-treatment interaction, kurtosis, and block-variance heterogeneity on the results obtained from the randomized blocks design are important papers in this section.

Many of the papers in section IV on testing statistical hypotheses, in particular those by Binder and Rozeboom on the logic of null hypothesis testing, demand careful study. Several of these papers are quite technical and probably too difficult for the typical graduate student in psychology or education. However, they deal with some extremely important matters related to the epistemology of statistical inference.

Section V, "Special Topics in Analysis of Variance," contains the largest number of papers of any section in the book. A reading of these papers should benefit almost all graduate students in behavioral science statistics courses. For example, the Millman and Glass paper on rules of thumb for writing the analysis of variance table

will be a blessing to anyone who has found that conventional statistics cookbooks do not contain the layouts for all possible experimental designs. In addition, the Evans and Anastasio paper on misuses of covariance analysis should be read by all who employ this method of controlling concomitant variables. Other important topics considered in this section are interaction, repeated measures designs, the homogeneity of variance assumption, and trend analysis with unequal n 's.

The final section of the book, "Correlation and Regression," should be of greatest interest to students and specialists in educational and psychological measurement. Following two papers on normality and other assumptions underlying the interpretation of the product-moment coefficient is a detailed, thirty-one page treatment by Darlington of problems and issues in multiple regression. A short paper on Bartlett's test of the significance of a correlation matrix, which will be a godsend to anyone who has vainly searched through Morrison (1967) or other books on multivariate statistical analysis for this type of test, completes the collection.

In sum, Heermann and Braskamp have put together in a single volume a series of important, well-written papers, many of which are inadequately handled in required courses in psychological and educational statistics. Of course, the reviewer could quibble with the editors at length about some of their choices, but after reading almost all of these papers I felt that my time had been very well spent.

REFERENCES

- Morrison, D. F. *Multivariate statistical methods*. New York: McGraw-Hill, 1967.
- Stevens, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology*. New York: Wiley, 1951.

LEWIS R. AIKEN, JR.
Guilford College

Benjamin Kleinmuntz (Ed.). *Clinical Information Processing by Computer: An Essay and Selected Readings*. New York: Holt, Rinehart and Winston, 1969. Pp. xi + 399. \$5.95 (paperback).

Benjamin Kleinmuntz is among the most knowledgeable and thoughtful workers in the field of personality measurement and whatever he writes in this area deserves careful attention. In the present work he has composed an essay which runs to 106 pages. For purposes of organizing his book, Kleinmuntz decided to break the essay into five sections and has inserted a number of readings into the gaps. Section I is titled, "Introduction: Computers as Computation and Noncomputational Information Machines." This is

followed by readings by Hovland, Green, and Newell and Simon. Section II is titled, "Personality Assessment: Computational Applications" and is followed by articles by Williams and Kleinmuntz and Eiduson et al. Section III is titled, "Personality Assessment: Noncomputational Applications of Computers" and is followed by readings by Tomkins, Iker and Harway, and Dunphy et al. Section IV is titled, "Medical Diagnosis: Computational and Noncomputational Applications" and is followed by readings by Lipkin et al., Lusted and Stahl, Entwisle and Entwisle, and Feurzeig et al. Finally, Section V sets forth the author's thoughts about the future for computer processing of clinical data.

Kleinmuntz's essay is clear and direct, albeit terse in a few spots. There is a notable lack of jargon without a sacrifice in ideas which should put the book within the reach of clinicians, the group for whom the book is mainly intended. To write with lucidity in a technical area is an achievement of no small proportion and Kleinmuntz has carried it off exceedingly well. The introductory section of the essay presents basic information about what computers are and how they operate in a general way, but one which is sufficient for the author's purpose. In the second section, Kleinmuntz sets forth his point of view:

Our point of view in this essay is a mechanistic one. Accordingly, we depict the psychologist as an information-processing organism who has collected direct observations, interviews, and tests as 'inputs' that he must process (analyze, organize and integrate) prior to 'outputting' his recommendations or predictions. (p. 85)

The author builds a strong and well documented case for his view that the computer is often as good if not better than humans in statistical processing of clinical information in psychodiagnostic work. The argument is vintage Meehl. The author is on weaker ground when, in Section III, he develops the case for the noncomputational uses of computers (an interesting anomaly but one that's been around for quite some time). Here, Kleinmuntz sees computers as playing an important role in the collection of information, e.g., mental status interviewing, according to some predetermined set of rules. He notes that the interview process contains sources of error arising from the *interviewer*, the *interviewee*, and the *interview* process. Kleinmuntz's hope, "... is that some of these sources of error will be minimized by automating the interview." (p. 150) Perhaps so, but what new sources of error will be introduced in its stead is impossible to estimate. Much work over a long period of time will be required to even begin to determine the possible roles that computers can play in the collection of psychological data.

In section IV, uses of computers in medical diagnosis, Kleinmuntz returns to solid ground. The basically Bayesian approach to computer use in this area is well explicated and several neat examples are presented. On the other hand, the discussion of the use of the computer as a teaching device in medical diagnosis seems more optimistic than the present state of development warrants. Kleinmuntz concludes his essay with a well balanced mixture of optimism and caution.

The injection of caution is interesting in its own right. When this writer was introduced to computers in 1961, the field was filled with an almost unbridled optimism. Newell and Simon pointed the way and many of us were quick to follow. Borko's *Computer Applications in the Behavioral Sciences* (1962) mapped out a brave new world. The future, however, did not run out as rosy as expected. After the initial flush of successes documented in Borko's book, there was a marked dry period in several areas and almost a total closing out of others, e.g., computer music. This is reflected in the publication dates of the articles Kleinmuntz selected for inclusion in his book. Of the thirteen readings, four were originally published before 1962, three in 1963, one in 1964, four in 1965 and one in 1966. The absence of more recently published works reflects, this writer believes, a notable lack of progress in the field rather than any lack of diligence on Kleinmuntz's part.

Where computer processing of clinical information has had its greatest successes and continues to be successful is in the area of statistical processing of data. Recent history has borne out the correctness of Meehl's position. Future progress, however, would seem to depend not on the development of new statistical procedures and computer software, but rather on the development of improved measuring instruments and procedures.

In summary, Kleinmuntz has produced a book intended mainly for clinicians that is stimulating and provocative albeit slightly overoptimistic. The deficiencies noted above are a reflection on the state of the field and not on the author.

RICHARD WOLF

Teachers College, Columbia University

Gérard Lemaire and Jean-Marie Lemaire (Eds.); *Psychologie sociale et expérimentation (Experimental procedures in social psychology)*; École Pratique des Hautes Études, Département des Publications, Paris: Mouton and Company, 1969, pp. 360, 28F (paperback).

This collection is directed toward beginning researchers and students in the area of social psychology. Twenty-one papers are grouped in five categories in the text. Each section focuses on a

particular area of experimental procedure. The identified categories are: general problems of psychological experiments, history and inherent difficulties of groups in experimental research, description of problems to be investigated, procedures and available instruments, social psychology of the experimental situation, and an introduction to simulation. The first four parts are preceded by an introduction which details the editors' position with respect to the authors' argument.

Eighteen of the twenty-one papers are translated from English. These are generally well known and available in this country (e.g., Donald T. Campbell's Factors relevant to validity of experiments in social settings, *Psychological Bulletin*, 1957, 54, pp. 297-312). Consequently, the book adds little to experimental methodology in the United States.

The collection and the original articles would be of considerable value to individuals in this country interested in developing their fluency in English-French translation.

ROBERT SMITH

University of Southern California

La recherche en enseignement programmé—tendances actuelles (Programmed learning research—major trends), actes d'un colloque O.T.A.N. Nice 1968, *Sciences du comportement* 8, collection dirigée par F. Bresson et M. de Montmollin, Paris: Dunod, 1969, pp. 360, 96F (paperback).

This collection of papers was presented at a symposium held in Nice, France, from May 13 to May 17, 1968. The meeting was sponsored by the Scientific Committee (Consultant Group of Human Factors) of the North Atlantic Treaty Organization (NATO). The symposium is one of a series that allows specialists from the several NATO countries to examine a subject of common interest.

The papers are organized under four general headings: Analysis and Structure of the Subject Matter, R. Gagné, co-chairman (United States); The Learning Process and Problem Solving, A. C. Atkinson, chairman (United States); Categories of Learning and Criteria for Evaluation of Learning Outcomes, R. Glaser, chairman (United States); Adaptive Machines, G. Pask, chairman (Great Britain). The proceedings reflect the same general preponderance of research from the United States. Programmed learning as viewed in this country vis-a-vis programmed teaching in the European countries dominated the discussions.

The general topics have received extensive consideration in this country and there seems to have been little of value added by moving the debates to a more salubrious clime.

ROBERT SMITH

University of Southern California

EDUCATIONAL and
PSYCHOLOGICAL



MEASUREMENT

Editor: W. Scott Gehman

Managing Editor: Geraldine R. Thomas

BOARD OF COOPERATING EDITORS

- DOROTHY C. ADKINS, *University of Hawaii*
LEWIS R. AIKEN, JR., *Guilford College*
HAROLD P. BECHTOLDT, *The University of Iowa*
WILLIAM V. CLEMANS, *Science Research Associates, Inc.*
LOUIS D. COHEN, *University of Florida*
JUNIUS A. DAVIS, *Educational Testing Service*
HAROLD A. EDGERTON, *Performance Research, Inc.*
MAX D. ENGELHART, *Duke University*
GENE V. GLASS, *University of Colorado*
E. B. GREENE, *Chrysler Corporation (Retired)*
J. P. GUILFORD, *University of Southern California, Los Angeles*
JOHN A. HORNADAY, *Babson College*
JOHN E. HORROCKS, *The Ohio State University*
CYRIL J. HOYT, *University of Minnesota*
MILTON D. JACOBSON, *University of Virginia*
JOSEPH C. JOHNSON II, *Duke University*
WILLIAM G. KATZENMEYER, *Duke University*
E. F. LINDQUIST, *State University of Iowa*
FREDERIC M. LORD, *Educational Testing Service*
ARDIE LUBIN, *Naval Medical Neuropsychiatric Research Unit, San Diego*
LOUIS L. MCQUITY, *University of Miami, Coral Gables*
WILLIAM B. MICHAEL, *University of Southern California, Los Angeles*
HOWARD G. MILLER, *North Carolina State University at Raleigh*
ELLIS B. PAGE, *The University of Connecticut*
NAMBU S. RAJU, *Science Research Associates, Inc.*
BEN H. ROMINE, JR., *University of North Carolina at Charlotte*
KENDON SMITH, *The University of North Carolina at Greensboro*
THELMA G. THURSTONE, *University of North Carolina at Chapel Hill*
HERBERT A. TOOPS, *The Ohio State University*
WILLARD G. WARRINGTON, *Michigan State University*
JOHN E. WILLIAMS, *Wake Forest University*
E. G. WILLIAMSON, *University of Minnesota*

VOLUME THIRTY-ONE, NUMBER TWO, SUMMER 1971

387
17.9.71
W

RELIABLE AND VALID HIERARCHICAL CLASSIFICATION^{1, 2}

LOUIS L. McQUITTY AND JEWEL M. FRARY

University of Miami
Coral Gables, Florida

SOME methods of hierarchical classification use only the highest index of association of every object with every other object; other methods use all indices of association (McQuitty, 1967, 1968; McQuitty and Clark, 1968). A problem is to use that particular set of indices of association which produces the most reliable and valid solution. This is what Reliable and Valid Hierarchical Classification attempts to accomplish.

Characterization of Types

Definitions

The method is derived from definitions of types.

Pure types. A *pure type* can be defined as a category of two or more objects of such a nature that every object in the category is more like every other object in the category, in terms of specified characteristics, than it is like any other object in any other category.

A *pure type* can also be defined as a category of two or more objects with a unique pattern of characteristics. Every object in the category possesses all of the characteristics of the pattern, and no object not in the category possesses all of them. The latter kind of object can possess some but not all of the characteristics.

¹This investigation was supported by Public Health Service Research Grant No. MH 14070-02 from National Institute of Mental Health.

²A revision and elaboration of a paper read at the annual meeting of the Society of Multivariate Experimental Psychologists, November 21, 1968, Austin, Texas.

If this latter definition of a *pure* type is translated into the same terminology as the first definition, it states that a type is a category of two or more objects of such a nature that every object of the category is identical, in terms of specified characteristics, with every other object of the category, and no object not in the category is identical with the objects of the category.

As another elaboration, a *pure* type can be defined as a category of two or more objects of such a nature that every object in the category is most like some other object in the category; every object is classified with the object most like itself.

These three definitions of a type are three specifications of interrelationships which characterize configurations constituting *pure* types. At the same time they suggest other configurations within the same general area as these but not necessarily satisfying completely the definition of any one of them. From this latter and more general point of view, a *pure* type is a category of two or more objects which belong together because of some inherent pattern of associations among the objects. There is no limit on the kind of configurations expressed by the patterns; they must, however, be held together by intrinsic associations.

Real types. *Pure* types exist in theory; their correlates in nature are called *real* types. They resemble *pure* types but usually do not conform completely to them.

A Numerical Display

Both *pure* and *real* types can be numerically displayed in matrices. One of the simpler ways of doing this is illustrated in the hypothetical data of Tables 1, 2, and 3.

The entry of 2 in Row C—Column A of Table 1, for example, reports that Object C is second most like Object A; Objects B and D are first and third most like Object A. The other entries are interpreted in an analogous fashion.

Table 1 illustrates fulfillment of the first *pure* type defined in this paper. In terms of the numerical display the two categories are called *square* types. The group of Objects A, B, C, and D, and the group of Objects W, X, Y, and Z, each constitute a *square* type. They are *pure* in the sense that no object has a rank larger than $n-1$ with any other object; n equals the number of objects in the submatrix. This condition means that every object in each

TABLE 1

Within-Column Rank Orders for a "Square" Type—Hypothetical Data

	A	B	C	D	W	X	Y	Z
A		1	2	2		Entries in this quadrant are from 4 to 7 inclusive.		
B	1		3	3				
C	2	2		1				
D	3	3	1					
W		Entries in this quadrant are from 4 to 7 inclusive.				2	2	1
X					3		1	2
Y					2	1		3
Z					1	3	3	

submatrix is more like every other object in that submatrix than it is like any object in any other submatrix.

In Table 2, Objects E, F, G, and H are identical and are different from Objects S, T, U, and V, which are also identical with one another. These two categories illustrate the second definition of a *pure type*, called an *identity type*.

Table 3 includes first another *square type*, Objects I and J, and secondly it illustrates the third definition of a type, represented in this case by an *elongated type*. Objects K and L are reciprocal; then Object L brings in Object M because Object M has Object L most like it. In a similar fashion, Object M brings in Object N, and Object N brings in Object O.

The essential associations of this latter type are shown more clearly in Figure 1, which portrays the elongated feature of the type.

Another configuration is labeled a *spotted type*; it resembles some one of the above configurations except that it varies from the standard configuration in certain cell entries and in specifiable amounts within those cell entries.

TABLE 2

Within-Column Rank Orders for an "Identity" Type—Hypothetical Data

Within-Column Rank Orders for an "Identity" Type—Hypothetical Data								
	E	F	G	H	S	T	U	V
E		1	1	1		Entries in this quadrant are from 4 to 7 inclusive.		
F	1		1	1				
G	1	1		1				
H	1	1	1					
S		Entries in this quadrant are from 4 to 7 inclusive.				1	1	1
T					1		1	1
U					1	1		
V					1	1	1	

TABLE 3

Within-Column Rank Orders for an Elongated Type—Hypothetical Data

	I	J	K	L	M	N	O
I		1					
J	1						
K				1			
L			1		1		
M						1	
N							1
O							

Note—All other entries are greater than one.

An example of a *spotted* type is taken from a hierarchical classification by Iterative, Intercolumnar Correlational Analysis (McQuitty and Clark, 1968) of real data and is reported in Table 4. Three types are reported in the table, one for each of the three Reciprocal Pairs 3-6 (or 3-16), 10-20, and 4-12. Objects 4 and 12, for example, are reciprocal because 4 is highest with 12 and 12 is highest with 4.

If each of these three types of 11, 6, and 3 objects respectively were a perfect representation of a pure, square type, then the first one would have no rank above 10, and the other two would have none above 5 and 2 respectively, i.e., none above $n-1$, where n = the number of objects in the type.

The footnoted values in the table indicate the ranks which are larger than the prescribed standard. This is the sense in which the types are spotted, and the size of the discrepant ranks shows how much they exceed the upper limit of $n-1$. The value, 14, in Row 13—Column 5, for example, has a deviation of $14-10 = 4$; $n-1 = 10$.

When data have been ranked within columns, a rank reflects a *spot* in a submatrix when its value is larger than $n-1$, where n is the number of objects in the submatrix and the ranks are those lifted from the matrix being divided.

Nonmembers

An advantage of the concept of *spots* is that it can be applied to indicate nonmembers. Object 8, classified with Objects 4 and 12 in

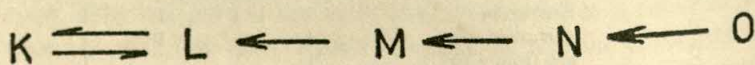


Figure 1. The essential structure of type KLMNO.

TABLE 4
Results of an Hierarchical Classification by Iterative, Intercolumnar Correlational Analysis*

	1	3	5	6	9	11	13	14	16	17	19	2	7	10	15	18	20	4	8	12
1																				
2		8	4	9	5	9	4	4	3	2	10									
3			2	1	1	3	5	1	1	1	3									
5		10		9	11*	6	15*	10	14*	10	5									
6		1	5		1	1	3	3	3	2	1									
8		2	7	2		4	1	4	2	8	4									
9		3									2									
10		6	1	3	5		9	8	7	13*	2									
11		9	7	6		7		6	3	6	8									
13		4	5	14*	4	8	5		3	5	8									
14		4	5	7	8	4	1	2	3	2	5									
16		1	1	5	11*	1		7	8		10									
17		4	4	5	10	10	7	6	8											
19	11*	6	2	3	5	2	10	10	9	13*										
2												2	3	5	2	7*	5			
7												4	1	2	7*	3	3			
10													6*	3	1	2	1			
15													8*	4	16*	4	4			
18													9*	4	2	1	2			
20												4	2	1		1	2			
3*																		3*	1	
10*																		10*	10*	
1																		1	5*	

* Ranks which are larger than the prescribed standard of $n - 1$, where n = the number of objects in the category.

the lower right-hand corner of Table 4, can be characterized as a nonmember because of the size of its deviation; out of the 20 objects of the original matrix it is only tenth most like each of the Objects 4 and 12 with which it is classified. In order to fit perfectly into the *square* type, it would have to be second most like Objects 4 and 12.

A problem is to determine when spots are sufficiently large to reflect nonmembers. All nonmembers yield spots but all spots do not reflect nonmembers. One practical approach is to require that every object be classified in every set of data; a spot then reflects a nonmember if, and only if, one or more objects are not correctly classified within the set of data. In an incorrect classification, there are some larger spots than there would be if all objects were correctly classified; they, therefore, reflect nonmembers. The correct classification of data which conforms perfectly to theory would eliminate all spots. The correct classification of data which does not conform would not eliminate all spots; it would, nevertheless, reduce at least some spots. Those spots which are eliminated or even merely reduced, in going from an incorrect to a correct classification, reflect nonmembers in the incorrect classification. In this approach, a correct classification must be determined by a criterion other than whether or not nonmembers are present; otherwise the definition of nonmembers and correct classification would both be circular. One such criterion is the logic of the method of classification.

Another way to characterize nonmembers is in terms of the data by which objects are classified; a spot reflects one or two nonmembers in a set of data if the one or two objects can be classified with a smaller deviation in another set of data.

In the latter approach, spots can be hypothesized to reflect nonmembers from two interrelated points of view: (1) the objects seem *not* to conform to the other objects of a set in terms of content, and (2) the nonfit is supported by the fact that the objects yield spots which are relatively large. Hypothesized nonmembers are then examined in a set of data with objects thought to be more appropriate to their classification and if they yield lower deviations in the new set of objects they are then confirmed as nonmembers in the original set of objects.

Nonmembers are generated also in a purely objective fashion.

If an analysis of n objects yields one or more types and one or more unclassified objects, the one or more unclassified objects are thereby nonmembers.

Every object is a nonmember at its terminal level of classification (from top down) where it stands alone and is contrasted only with other nonmembers. In order to distinguish these latter nonmembers from those characterized above, they are called entities rather than nonmembers.

When a matrix divides to yield one nonmember and a submatrix of two or more other objects, the latter submatrix constitutes a pseudotype. In order to fulfill the requirement of a type, a submatrix must be separated from at least two other objects.

If every division throughout a hierarchical analysis yields the above kind of result, then the entire structure is pseudotypological.

Mental Health Theory

The concept of *spotted types* has another advantage. It is hypothesized that "mental patients" reflect more nonmembers, more *spots*, and more extreme *spots* than "normals" and that they do this more extensively in matrices of interrelationships between characteristics within the single individual than they do in matrices reporting relationships between persons. These are hypotheses for later studies; their investigation is made possible by the approaches of this paper.

The Method

The concept of *pure types* can be used to generate a simple method for the isolation of *real types*.

Hypothetical Data

The method is generated with specific reference to the first definition of pure types and with hypothetical data which fulfill the definition.

Tables 5 and 6 report rank orders within columns for the members of two sets of pure types; Objects B, C, and D of Table 5, for example, are first, second, and third most like Object A. Other column entries are interpreted in an analogous fashion.

Table 7 combines in a single matrix the within column ranks of

TABLE 5

Within-Column Rank Orders for the First Set of Hypothetical Data

	A	B	C	D
A		1	2	2
B	1		3	3
C	2	2		1
D	3	3	1	

the first two tables. In this larger matrix, the objects of the first two tables are intermingled.

By definition of a pure type, all of the empty cells of Table 7 (other than the diagonals) have entries larger than $n-1 = 3$. This is because the two square types are required by definition to have all of the entries of $n-1$ and smaller.

The two types reveal themselves clearly in the patterns of ranks of Table 7 as encompassed by the lines enclosing each of the two types.

There is, however, another simple way for isolating the types, even if the types are only *real*—not *pure*—and do not reveal themselves clearly in a form such as illustrated in the table.

In the two columns on the extreme right of Table 7 are reported first the number of ranks of one in each row and then the number of ranks of one and two combined in each row. The numbers in the first of these two columns are the criteria for classifications in terms of a rank of one, and those in the second column are criteria in terms of ranks of one and two combined.

The analysis starts with the highest rank (one) and the largest criterion for one and proceeds first by lowering the criterion for ranks of one and next by lowering the ranks to include one and two; one, two, and three; one, two, three, and four, etc. Within each decrease in rank the largest criterion is first used and then

TABLE 6

Within-Column Rank Orders for the Second Set of Hypothetical Data

	W	X	Y	Z
W		2	2	1
X	3		1	2
Y	2	1		3
Z	1	3	3	

TABLE 7

Within-Column Ranks of Tables 5 and 6 Combined in a Single Matrix

									Number of Ranks of	
	A	W	X	C	D	B	Y	Z	One	One and Two
A				2	2	1			1	3
W			2				2	1	1	3
X		3					1	2	1	2
C	2				1	2			1	3
D	3			1		3			1	1
B	1			3	3				1	1
Y		2	1					3	1	2
Z		1	3				3		1	1

successively lower criteria until exhaustion before proceeding to a further decrease in rank.

Every time the ranks are lowered, such as from one to one and two combined, the classification starts over again.

The largest criterion for a rank of one in Table 7 is one, and it is reported in all eight rows. The starting point for tied criteria is unimportant, but all of them must, of course, be used.

Row A of Table 7 assigns Objects A and B to a common type because B has a rank of one in this row. This fact is summarized in the first row of Table 8 by an asterisk in Row A and Column B. Analogously, Row B of Table 7 assigns Objects A and B to the same common type as reported in the sixth row of Table 8.

Neither Object A nor B, nor any other object, assigns any other member to the above type. These two members constitute a type; A is most like B, and B is most like A. In an analogous fashion, the following pairs of objects constitute types: CD, WZ, and XY.

Table 7 is further analyzed in Table 9 to yield two types of four objects each. The last Column of Table 7 is utilized in this effort; it reports the number of ranks of one and two combined in each

TABLE 8

Assigning the Objects of Table 7 to Types of Two Objects Each

A	W	X	C	D	B	Y	Z	Criterion
A					*			1
	W						*	1
		X				*		1
			C	*				1
			*	D				1
*					B			1
		*				Y		1
	*						Z	1

row of Table 7. These entries are the criterion values for the analysis. In analyzing the last Column of Table 7, the initial step applies to the row or rows with the largest criterion, three in this case. There are three rows with a criterion of three. The start is with the first criterion of three, from top down, chosen arbitrarily. Under the criterion of three, Row A of Table 7 assigns Objects A, B, C, and D to a type, as summarized in Table 9 by asterisks in Row A and Column C, D, and B; Row W assigns Objects W, X, Y, and Z to a type; and Row C assigns A, B, C, and D to a type, confirming the action of Row A.

The analysis is not carried beyond the above point because the matrix of Table 7 has now been bifurcated into two submatrices in terms of the most dependable indices, those reflecting ranks of one and two and excluding the larger ranks of three and above. In this case, continuing with the larger and less dependable ranks would reconfirm the above results. Such is not, however, necessarily the case when isolating *real* types rather than *pure* types.

The above analysis was initiated by using first the most dependable ranks, viz., those of one which are based on the highest index of association within every column. The analysis proceeded toward bringing in the lower ranks, first the ranks of two, then three, four,

TABLE 9

Assigning the Objects of Table 7 to Types of Four Objects Each

A	W	X	C	D	B	Y	Z	Criterion
A			*	*	*			3
	W	*				*	*	3
*			C	*	*			3

etc. until the matrix was bifurcated. Ranks of one and two bifurcated the matrix of the present example.

Real Data

Table 10 reports agreement scores between spoons based on the presence and absence of characteristics as judged by a single subject.

The data were chosen for illustrating the present method because they contain a number of unique problems and they had proven difficult to analyze in an earlier study (McQuitty, Price, and Clark, 1967).

Table 11 reports the ranks within columns for the data of Table 10. In the case of a tie in agreement scores, all tied scores are assigned the rank which would be given if there were only one score at that value and all of the other scores were smaller. For example, the seven highest agreement scores in Column 16 of Table 10 are 34, 33, 30, 30, 30, 30, and 29. They are assigned ranks of 1, 2, 3, 3, 3, 3, and 7 respectively, as reported in Column 16 of Table 11.

Column 2 of Table 12 reports the number of ranks of one in every row of Table 11. Successive columns of Table 12 report the number of ranks of 1 and 2; 1, 2, and 3; and 1, 2, 3, and 4, respectively, in every row of Table 11.

Columns 2, 3, 4, and 5 of Table 12 are used to assign spoons to types. The highest criterion in Column 2 of Table 12 is 5. It pertains to Spoon 3. The further analysis begins, therefore, with Spoon 3. As shown by the ranks of 1 in Row 3 of Table 11, Spoon 3 assigns itself and five other spoons to Type 1, viz., 3, 6, 9, 14, 16, and 17. This fact is reported in Table 13.

The next highest criterion is 4 for Spoons 6 and 16, as shown in Table 12. Since there is a tie, both spoons will be analyzed; the one which is selected first is immaterial. Spoon 6 assigns Spoons 3, 6, 9, 11, and 19 to a type, determined to be Type 1 because of the overlap in assignments of these spoons with spoons already assigned to Type 1. Spoon 16 assigns Spoons 16, 1, 3, 9, and 13 to Type 1.

The analysis continues by using assignment spoons with successively smaller criteria until all spoons are assigned. Spoon 10, under a criterion of 3, assigns Spoons 10, 7, 15, and 20 to a type, determined tentatively to be Type 2 because there is no overlap

TABLE 10
*Agreement Scores between Spoons**

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		20	29	20	25	24	20	13	29	20	23	18	28	28	24	30	28	16	22	20
2	20		20	25	17	15	26	20	15	25	13	26	20	20	27	21	25	20	18	25
3	29	20		19	26	34	23	13	33	20	30	18	27	31	26	34	32	19	30	22
4	20	25	19		22	18	25	20	18	22	18	27	20	19	25	19	22	21	19	25
5	25	17	26	22		24	17	23	23	16	28	20	18	23	20	21	24	14	26	14
6	24	15	34	18	24		20	12	33	17	32	15	29	29	23	30	28	16	32	20
7	20	26	23	25	17	20		14	19	30	18	26	18	22	25	23	26	24	16	28
8	13	20	13	20	23	12	14		12	15	16	18	13	11	21	11	13	15	18	12
9	29	15	33	18	23	33	19	12		20	29	15	30	28	26	33	25	17	29	21
10	20	25	20	22	16	17	30	15	20		14	20	21	21	28	22	22	27	15	31
11	23	13	30	18	28	32	18	16	29	14		16	15	24	21	29	22	13	31	16
12	18	26	18	27	20	15	26	18	15	20	16		15	17	23	15	19	21	18	21
13	28	20	27	20	18	29	18	13	30	21	25	15		27	26	30	26	17	23	23
14	28	20	31	19	23	29	22	11	28	21	24	17	27		24	30	27	19	23	22
15	24	27	26	25	20	23	25	21	26	28	21	23	26	24		25	24	22	24	27
16	30	21	34	19	21	30	23	11	33	22	29	15	30	30	25		28	17	26	23
17	28	25	32	22	24	28	26	13	25	24	22	19	26	27	24	28		18	22	22
18	16	20	19	21	14	16	24	15	17	27	13	21	17	19	22	17	18		14	30
19	22	18	30	19	26	32	16	18	29	15	31	18	23	23	24	26	22	14		16
20	20	25	22	25	14	20	28	12	21	31	16	21	23	22	27	23	22	30		

* From McQuitty, Price and Clark, 1907.

TABLE 11
*The Indices of Table 10 Converted to Ranks Within Columns**

	Code Number of Spoons																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	9	8	10	4	9	12	11	5	11	9	10	4	4	10	3	2	14	10	14	1
2	12	14	2	15	17	3	3	17	5	18	2	13	15	2	14	8	7	13	5	2
3	2	9	13	2	1	9	11	1	11	3	10	5	1	4	1	1	8	3	9	3
4	12	4	16	10	14	6	3	15	7	12	1	13	16	7	16	13	5	12	5	4
5	7	16	10	6	9	17	1	11	16	6	7	15	10	19	14	10	17	5	18	5
6	8	17	1	17	5	12	15	1	15	1	16	3	3	14	3	2	14	1	14	6
7	12	2	12	2	15	12	10	14	2	12	2	15	12	7	11	6	3	16	3	7
8	19	9	19	10	7	19	19	19	17	14	10	19	19	17	19	19	16	13	19	8
9	2	17	3	17	7	2	14	15	11	17	4	1	4	4	2	8	11	4	12	9
10	12	4	14	6	17	15	1	8	13	19	7	12	14	1	13	10	2	18	1	10
11	10	19	6	17	1	3	15	7	5	17	15	9	8	17	7	13	19	2	16	11
12	17	2	18	1	12	17	3	5	17	11	14	18	6	4	18	17	5	13	12	12
13	4	9	9	10	14	6	15	11	4	9	7	5	5	4	3	6	11	8	7	13
14	4	9	5	13	7	6	11	18	8	9	8	14	7	10	3	5	8	8	9	14
15	8	1	10	2	12	11	6	2	9	11	4	4	7	8	10	10	4	7	4	15
16	1	8	1	13	11	5	9	18	1	7	16	1	1	2	7	2	11	5	7	16
17	4	4	4	6	5	8	3	11	10	6	10	9	7	6	8	8	10	10	9	17
18	18	9	16	9	18	16	8	8	16	4	18	5	17	16	16	17	18	19	2	18
19	11	15	6	13	2	3	18	5	5	17	2	10	10	10	10	9	13	17	16	19
20	12	4	13	2	18	12	2	15	12	1	14	5	10	12	2	11	13	1	16	20

* The highest rank (smallest number) is used in case of ties; ranks 1, 2, 2, and 5 would be assigned to indices of 34, 32, 32, and 30.

TABLE 12

Criterion Values for the Rows of Table 11

Code Number of Spoons	Number of ranks of :			
	1	1 and 2	1, 2, and 3	1, 2, 3, and 4
1	0	1	2	5
2	0	3	5	5
3	5	7	9	10
4	1	1	2	3
5	1	1	1	1
6	4	5	8	8
7	0	4	6	6
8	0	0	0	0
9	1	4	5	9
10	3	4	4	5
11	1	2	3	3
12	1	2	3	3
13	0	0	1	4
14	0	0	1	2
15	1	3	4	7
16	4	6	6	7
17	0	0	1	4
18	0	1	1	2
19	0	2	3	3
20	2	5	5	6

with Type 1. Spoon 20, under a criterion of 2, assigns Spoons 20, 10, and 18 to Type 2, so determined because of the overlap of these spoons with those already assigned to Type 2. Spoon 4, under a criterion of one, assigns Spoons 4 and 12 to a new type, tentatively Type 3, because there is no overlap with previous types.

Analogously, under a criterion of one, Spoon 5 assigns itself and Spoon 8 to Tentative Type 4; 9 assigns itself and 13 to Type 1; 11, itself and 5 to Type 1; 12, itself and 4 to Type 3; and 15, itself and 2 to Type 2.

For convenience in reading the typal memberships, they are summarized in the bottom row of Table 13. There is but one conflict in the above assignments; Spoon 5 with a criterion of one assigned itself and Spoon 8 to Type 4, but also with a criterion of one, Spoon 11 assigned itself and Spoon 5 to Type 1. This latter action also assigned Tentative Type 4 to Type 1 and thus eliminated Tentative Type 4. This occurred because once Spoon 5 was assigned to Type 1, Tentative Type 4 produced an overlap with Type 1. The other two types, 2 and 3, are defined without conflict.

A complete hierarchical analysis often classifies the objects

TABLE 13

Assigning Spoons to Types Based on Ranks of One

Types	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Criterion
1			3			*			*					*		*	*				5
1			*			6			*		*								*		4
1	*		*						*				*			16					4
2							*			10					*					*	3
2										*				*				*	20		2
3				4								*									1
4					5			*													1
1									9				*								1
1					*						11										1
3			*									12									1
2	*														15						1
Types	1	2	1	3	4, 1	1	2	4, 1	1	2	1	3	1	1	2	1	1	2	1	2	

into two categories at the top level of classification. In the above results there are three categories. The analysis is continued using ranks of 1 and 2 combined in lieu of just a rank of 1.

Table 14 derives from Table 11 and from Column 3 of Table 12 for ranks of 1 and 2 combined in the same fashion that Table 13 was derived from these tables for ranks of 1, exclusively.

By the time the criterion of 3 for ranks of 1 and 2 combined (Column 3, Table 12) is exhausted, all spoons are assigned to Types 1 and 2 in Table 14, and there are no conflicts. The analysis of the original matrix into two categories is complete.

In order to illustrate the eventual degeneration of assignments in relation to declining indices of reliability and validity as lower ranks are incorporated into the analysis, the method was applied to ranks of 1, 2, and 3 combined and to ranks of 1, 2, 3, and 4 combined, with the results reported in Tables 15 and 16 respectively.

The classification into types by ranks of 1, 2, and 3 combined confirmed the classification by ranks of 1 and 2 combined. Ranks of 1, 2, 3, and 4 failed, however, to yield a classification of all objects into one of two types.

As summarized in Table 16, criteria of 10, 9, and 8 assigned all objects except 2, 4, 7, 8, 10, 12, 18, and 20 to Type 1. A criterion of 7 had to be applied. Using first Object 15, with a criterion of 7, all objects, except Object 7, were assigned to Type 1. Continuing with a criterion of 7 and using Object 16, Object 7 still remained unassigned with all other objects assigned to Type 1.

TABLE 14
Assigning Spoons to Types Based on Ranks of One and Two Combined

Types	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Criterion
1	*		3	*	*	*			*					*		*	*				7
1	*		*						*				*	*		16	*				6
1			*			6			*		*						*		*		5
2				*			*			*		*			*			*		20	5
2		*		*			7			*		*									4
1	*					*			9	10			*			*		*		*	4
2							*					*			*						4
2		2		*				*							*						3
2		*		*											*						3
Types	1	2	1	2	1	1	2	2	1	2	1	2	1	1	2	1	1	2	1	2	

TABLE 15
Assigning Spoons to Types Based on Ranks of One, Two, and Three Combined

Types	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Criterion
1	*		3		*	*			*		*		*			*	*		*		9
1			*			6		*	*		*		*			*	*		*		8
2		*		*			7		*	*		*					*		*		6
1	*		*					*					*			16	*				6
2		2		*			*	*				*		*		*					5
1	*		*			*		9				*				*					5
2				*			*		*	*					*		*	*		20	5
Types	1	2	1	2	1	1	2	2	1	2	1	2	1	1	2	1	1	2	1	2	2

TABLE 16
Assigning Spoons to Types Based on Ranks of One, Two, Three, and Four Combined

[illegible]

Proceeding with a criterion of 6, all objects were assigned to Type 1. Consequently, the ranks of 1, 2, 3, and 4 combined failed to assign every object to one of two types.

The classification by ranks of 1 and 2 combined (and confirmed by 1, 2, and 3 combined) is favored over that of 1, 2, 3, and 4 combined because it is based on higher ranks and it classifies every object into one of two classes.

The indication of the above results is that the relatively high indices of association between objects can be used to yield reliable and valid assignment to types and that the lesser indices sometimes fail and cannot always be used with the same dependability.

The analysis is continued by applying the above procedures separately to each of the two types (or submatrices) which derived from the above analysis, using ranks of either 1 and 2 combined, or 1, 2, and 3 combined.

The procedures are applied here to the analysis of the larger of the two submatrices to illustrate some results not obtained in the above analysis.

Table 17 reports the ranks within columns of the agreement scores of the spoons of Type 1, as isolated in Table 14. A column on the right reports the ranks of one in every row of the table. The information of this column is used to assign the spoons to types as shown in Table 18. Spoon 5 drops out because *all* other spoons are assigned prior to it. The other spoons constitute a pseudotype. Spoon 5 is a nonmember at this level of classification.

A result like the above is a solution because it fulfills our definition of a pseudotype. It does not, however, classify all objects into one of two types which is a goal to be realized when possible.

Whether or not the latter purpose can be realized is determined by continuing the steps of the analysis as summarized in Tables 19, 20, and 21.

The additional analysis shows that the best solution by ranks of 1 and 2 combined is to assign all but two objects to a type. The best solution by ranks of 1, 2, and 3 combined and also by ranks of 1, 2, 3, and 4 combined is to assign all but one object to a type, viz., Object 13. Ranks of 1, 2, 3, 4, and 5 combined yield a criterion of 10 (out of a total of 11 objects) and therefore assign all objects a single category (composed of the assignment object and the other 10 objects). Increasing the ranks will continue to yield

TABLE 17
Ranks Within Columns of Agreement Scores for the Spoons of Type 1

Code Numbers of Spoons	Number of Ranks of:															
	1	3	5	6	9	11	13	14	16	17	19	1	1 & 2	1, 2, & 3	1, 2, 3, & 4	1, 2, 3, 4, & 5
1		8	4	9	5	9	4	4	3	2	9	0	1	2	5	6
3	2	10	2	1	1	3	5	1	1	1	3	5	7	9	9	10
5	7		9	9	10	6	10	9	10	8	5	0	0	0	0	1
6	8	1	5	2	1	1	3	3	3	2	1	4	5	8	8	9
9	2	3	7			4	1	4	2	7	4	1	4	5	8	8
11	9	6	1	3	5		8	8	7	9	2	1	2	3	3	4
13	4	9	10	6	4	7	6	6	3	6	7	0	0	1	3	3
14	4	5	7	6	8	8	5		3	5	7	0	0	1	2	5
16	1	1	9	5	1	4	1	2	3	2	5	4	6	6	7	9
17	4	4	5	8	9	10	7	6	8	2	9	0	0	0	2	3
19	10	6	2	3	5	2	9	9	9	9		0	2	3	3	4

TABLE 18

Assigning Spoons of Table 17 to Types Based on Ranks of One

Type	1	3	5	6	9	11	13	14	16	17	19	Criterion
1		3		*	*			*	*	*		5
1		*		6	*	*					*	4
1	*	*			*		*		16			4

the same result because the criterion never decreases with an increase in ranks.

The original solution based on the rank of one is accepted because it classifies as many objects as any other solution and does it on the basis of higher ranks.

Table 22 reports the ranks within columns of the agreement scores of the pseudotype. A rank of 1 with successive criteria first of 5 and then of 4 yields the results shown in Table 23. The criterion of 5 fails to assign the objects to two types. Criteria of 4 and 5 place all objects in the same category.

Ranks of 1 and 2 with successive criteria first of 6 and then of 5 and 6 yield the results shown in Table 24. A criterion of 6 fails to assign all of the objects and the criteria of 5 and 6 assigns all objects to the same category.

In situations like the above (represented by Tables 23 and 24 and also earlier in Table 17) there are at least two possible solutions. One solution involves an assumption: if a rank of x assigns all objects to a single category without realizing a solution, it is assumed that every larger rank would also assign them all to a single category without realizing a solution. Under this approach, the investigator returns to an earlier stage of the analysis which did not assign every object to but one category. He selects that stage which classifies the most objects. It is in this case a rank of 1 and 2 and a criterion of 6 (Table 24). It assigned Objects 1, 3, 6, 9, 13, 14, 16, and 17 to a type and left Objects 11 and 19 un-

TABLE 19

Assigning Spoons of Table 17 to Types Based on Ranks of One and Two Combined

Type	1	3	5	6	9	11	13	14	16	17	19	Criterion
1	*	3	*	*	*			*	*	*		7
1	*	*			*		*	*	16	*		6
1		*		6	*	*				*	*	5

TABLE 20

Assigning Spoons of Table 17 to Types Based on Ranks of One, Two, and Three Combined

Type	1	3	5	6	9	11	13	14	16	17	19	Criterion
1	*	3	*	*	*	*	*	*	*	*	*	9
1		*		6	*	*	*	*	*	*	*	8

assigned and therefore nonmembers. (The approach could have yielded more than one type and one or more nonmembers.)

An alternative approach to granting the above assumption is to investigate it, i.e., to proceed with larger numerical ranks. The combined ranks of 1, 2, and 3 with the top criterion for them of 8 assign all objects to a single category as shown in Table 25. Ranks of 1, 2, 3, and 4 combined act in the same fashion. Ranks of 1, 2, 3, 4, and 5 combined yield a criterion of 9 and necessarily assign all 10 cases (the assignment object and the 9 associates) to a single category. Any further ranks would do likewise because they too would yield a criterion of nine. In this particular set of data, the assumption is substantiated.

When the above alternatives do not agree, the alternative which classifies the most objects into types is accepted as the better.

The analysis proceeds in the above fashion until all objects have been assigned to types (which cannot be further decomposed by the method) and nonmembers. If more than two nonmembers are generated throughout the entire analysis, they are assembled into a matrix and the entire method is applied to them. The results of the complete analysis are shown in Figure 2.

Results

The first level of classification assigned all 20 objects to one of two types. The second level generated one nonmember (Object 5) and assigned the other 19 objects to one of three types. The third

TABLE 21

Assigning Spoons of Table 17 to Types Based on Ranks of One, Two, Three, and Four Combined

Type	1	3	5	6	9	11	13	14	16	17	19	Criterion
1	*	3	*	*	*	*	*	*	*	*	*	9
1		*		6	*	*	*	*	*	*	*	8
1	*	*		*	9	*	*	*	*	*	*	8

TABLE 22

Ranks Within Columns Derived from Agreement Scores

											Number of Ranks of:			
	1	3	6	9	11	13	14	16	17	19	1	1 & 2	1, 2, & 3	
1			8	9	5	8	4	4	3	2	8	0	1	2
3		2		1	1	3	5	1	1	1	3	5	6	8
6		7	1		1	3	3	3	2	1	4	5	5	8
9		2	3	2		4	1	4	2	7	4	1	4	5
11		8	6	3	5		8	8	7	8	2	0	1	2
13		4	9	6	4	6		6	3	6	6	0	0	1
14		4	5	6	8	7	5		3	5	6	0	0	1
16		1	1	5	1	4	1	2		2	5	4	6	6
17		4	4	8	9	9	7	6	8		8	0	0	0
19		9	6	3	5	2	9	9	9	8		0	1	2

level generated four nonmembers and two entities and assigned the other 13 objects to one of two types. The fourth level generated three nonmembers and assigned the other 10 objects to types. The fifth level of classification generated 10 entities; the two submatrices from which they came did not generate other types.

Summary of Procedure

A matrix of interassociations between objects is prepared. The interassociations are ranked within columns, with the largest association being assigned a rank of one, the next largest a rank of two, the next largest a rank of three, etc. In the case of a tie such as 34, 33, 30, 30, 30, 30, and 29, the ranks would be 1, 2, 3, 3, 3, 3, and 7 respectively; the highest rank (smallest numerical value) represented by the tied values is assigned to all of them, and the rank of the next lower association is based on the assumption that the tied values used up as many ranks as there are tied values. These steps yield a matrix of ranks within columns.

The next step is to count the number of ranks of one within each row of the matrix and record it in a column labeled criterion for

TABLE 23

Assigning Spoons of Table 22 to Types Based on Ranks of One

Type	1	3	6	9	11	13	14	16	17	19	Criterion
1		3	*	*			*	*	*		5
1		*	6	*	*					*	4
1	*	*		*		*		16			4

TABLE 24

Assigning Spoons of Table 22 to Types Based on Ranks of One and Two

Type	1	3	6	9	11	13	14	16	17	19	Criterion
1	*	3	*	*			*	*	*		6
1	*	*		*		*	*	16	*		6
1		*	6	*	*				*	*	5

ranks of one; if a row contains five ranks of one, then it has a criterion of five for ranks of one.

An effort is made to assign all objects to types on the basis of the criteria for ranks of one. The largest criterion is used first. If more than one object (row) report the largest criterion, then all objects tied for the top criterion must be used one at a time in attempting the assignments of objects to types, and it is immaterial which object is used first, second, third, etc. An object with the highest criterion assigns itself and all objects with a rank of one in its row; they are assigned tentatively to Type 1.

Assume that the object did not assign all other objects. Another object with the highest criterion (if there are ties) or an object with the next highest criterion (if there are no ties for the highest) assigns objects to a type based on a criterion of one. If there are overlaps between the objects of this type and Type 1, then all of the assignments are to Type 1. If there are no overlaps, the new assignments are to Type 2.

The above steps complete the classification at the first level of assignments if and only if every object is assigned to one of two types and there is not another object tied in criterion with the object just used in making assignments.

If the above steps do not complete the first level of classification, another object with the highest criterion is chosen (if there are three or more ties for the highest criterion) or an object with the second highest criterion (if there are two objects for either the first or second highest criterion) or an object with the third highest

TABLE 25

Assigning Spoons of Table 22 to Types Based on Ranks of 1, 2, and 3

Type	1	3	6	9	11	13	14	16	17	19	Criterion
1	*	3	*	*	*		*	*	*	*	8
1		*	6	*	*	*	*	*	*	*	8

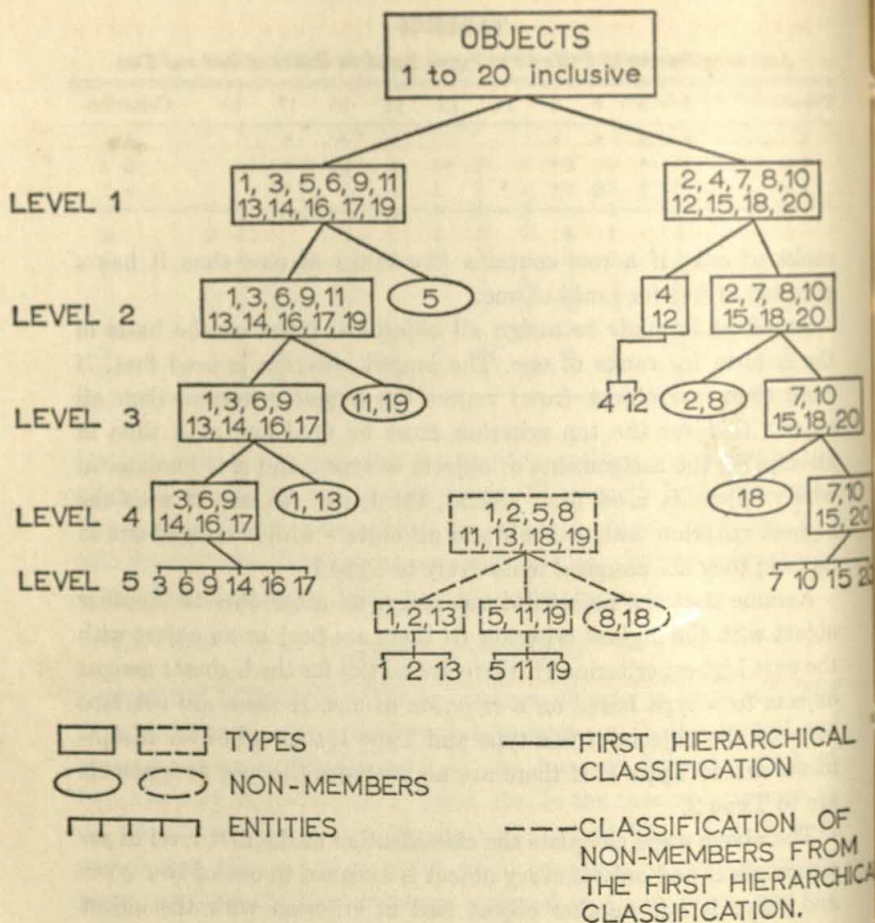


Figure 2. Reliable and valid hierarchical classification.

criterion (if there are no ties for either first or second highest criterion). It assigns itself and all objects with a rank of one in its row to a type. If the new type has one or more overlaps in assignments with but one other type, it becomes a part of the type with which it has overlap.

If the new type has overlap with more than one type, it and all of the types with which it has overlap are combined into a single type.

The above steps continue until all objects have been assigned to either one of two types or to one type.

In the case of the latter outcome, the rank on which the criteria is based is increased by one and the count of the number of qualifying ranks within rows is the number of ranks of x (2 in the present stage) and numerically less.

The process repeats the above steps until either all objects are assigned to one of two types or a criterion of $m-1$ has been reached, where m is the number of objects in the matrix being subdivided; a criterion of $m-1$ assigns its object and the $m-1$ other objects to a single type.

When a criterion of $m-1$ has been reached, the process reverts to the rank and criterion which classified most objects into types (preferably without classifying all of them into one category); objects not classified are categorized as *nonmembers*. If there is a tie for the minimum number of nonmembers, the classification by the higher ranks is chosen. The method cannot use one object with a criterion of x for ranks of y and higher without using all objects with a criterion of x for ranks of y and higher.

When the classification of the original matrix has been completed, the steps are repeated with the submatrices (types and pseudotypes) derived from the original matrix and then the successive submatrices, until at the bottom level every object stands alone as an entity.

All nonmembers produced at other than their respective bottom levels of classification (i.e., all nonmembers which are not entities) are assembled into a matrix and the process is repeated on them.

If in the analysis of nonmembers, more than two nonmembers are again generated, the process is repeated until all possible nonmembers are classified.

Summary

Pure types exist only in theory and are free from characteristics which do not conform to typical memberships. *Real* types, on the other hand, exist in nature; they approach but frequently do not correspond perfectly to *pure* types.

Insofar as *real* types do not conform to *pure* types, they can be characterized as *spotted*. Both the number of *spots* and the extent of discrepancy of each *spot* from its *pure* correlate can be determined.

"Mental patients" probably reflect more *spots* than "normals," and their *spots* are probably more discrepant than those of "nor-

mals." "Mental patients" probably reflect these differences more clearly in matrices reporting interassociations between characteristics of the single individual than in matrices reporting interassociations between persons.

In a matrix of interassociations between the members of several *pure* types, every index of interassociation reflects valid typal membership. In a matrix of interassociation between members of several *real* types, on the other hand, only the higher indices generally reflect typal relationships validly. Consequently, in the isolation of *real* types, by the method of this paper, only the larger indices of interassociations are used. More specifically, the indices are assigned rank orders within columns of a matrix. A solution is attempted first in terms of ranks of one exclusively, then one and two exclusively, then one, two, and three exclusively, etc. The analysis at each level of hierarchical classification concludes with the earliest and, thus, the most valid solution. However, a continuance until conflicts arise gives additional objective evidence related to validity and further objective evidence of the similarity of the *real* types to *pure* types.

REFERENCES

- McQuitty, L. L. A mutual development of some typological theories and pattern-analytic methods. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 21-46.
- McQuitty, L. L. Improving the validity of crucial decisions in pattern analytic methods. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 9-21.
- McQuitty, L. L. and Clark, J. A. Clusters from iterative, inter-columnar correlational analysis. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 211-238.
- McQuitty, L. L., Price, L., and Clark, J. A. The problem of ties in a pattern-analytic method. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 787-796.

SYSTEMATIC SCORING OF RANKED DISTRACTORS FOR THE ASSESSMENT OF PIAGETIAN REASONING LEVELS¹

DAVID H. FELDMAN² AND WINSTON MARKWALDER

University of Minnesota

GUTTMAN and Schlesinger (1967) recommended that item distractors be systematically constructed to increase the information yield of test devices. In describing the typical purpose of distractors Guttman and Schlesinger wrote: "Keeping the correct answer company is usually regarded as the only function of distractors, and sufficient attraction is deemed sufficient qualification for being a good distractor" (p. 569). At least three desirable additional benefits can be derived from distractors, according to Guttman and Schlesinger:

1. Successful prediction of relative empirical difficulties of distractors
2. Reduction of variation in test results due to undesired factors
3. Possibility of differential scoring of subjects on the types of wrong answers to which they are attracted.

The research reported by Guttman and Schlesinger (1967) provides data bearing on the first two propositions; the research reported in this paper bears on the third. Systematic scoring of ranked dis-

¹ The research reported here was supported in part by grants from the Stanford Center for Research and Development in Teaching (OE-6-10-078) and the Research and Development Center in the Education of Handicapped Children, University of Minnesota (OEG-0-9-332189-4533-032). The authors acknowledge the helpful assistance of Katherine Gray and Lee Ellen Johnson, and thank the principals and teachers of the schools in San Francisco, California and St. Paul, Minnesota who participated in the study. This report is based on a presentation given at the American Educational Research Association annual meeting, Minneapolis, March 1970.

² Reprints may be requested from the first author, Pattee Hall, University of Minnesota, 55455. Copies of the Map Test are available from the first author.

tractors was used in this study to diagnose types of wrong answers as they reflect different Piagetian reasoning levels. Specifically, the purposes of the study were threefold: first, to determine if a test of achievement in a specific content area could also be used to diagnose the level of thinking at which a child tends to respond; second, to compare a "reasoning level" analysis with a more conventional index of performance; and third, to test an assumption of Piagetian theory as it pertains to a specific school task. The assumption is that reasoning development is attained in a fixed sequence of stages by all children, although not necessarily at the same rate.

A new spatial reasoning (map reading) task was designed to gather data on the efficacy of a ranked distractor instrument for educational diagnosis. The task was based on a conceptual analysis of the geographic map which attempted to specify the skills requisite to proper map understanding (Salomon, 1968; Salomon and Feldman, 1969); the technique was somewhat similar to that used by Gagne' (1962) for analysis of skills requisite to long division. The test consisted of 25 items constructed to assess skills requisite to map drawing; the items designed to measure these skills were ordered into a hypothesized fixed sequence of acquisition. Previous studies (Feldman, 1969, 1970) indicated that performance on these 25 items was positively related ($r = .46$, $df = 268$, $p < .01$) to a map-drawing criterion, thus giving some measure of concurrent validity to the map-reading test. Other validation techniques adapted from Eisner (1967) and Piaget and Inhelder (1967) are reported in detail elsewhere (Feldman, 1969, 1970).

Of the 25 test items, 17 were multiple-choice with four distractors and a blank space if *S* wished to write his own answer. The remaining eight items required *Ss* to perform some reasoning operation (such as indicating the four directions around a map) or to write an answer (such as a rationale for picking a city as the capitol of a mapped island). All 25 items were designed to induce responses at four reasoning levels based on Piaget's theory of cognitive development (Piaget, 1950; Piaget and Inhelder, 1967; Sullivan, 1967). The reasoning levels were selected to correspond to the major stages between six and 14 years. In the case of the multiple choice items, each of four distractors was designed to reflect (a) tautological or imaginative reasoning, (b) perceptual/associative reasoning, (c) concrete reasoning, or (d) formal reasoning. For the remaining eight

Items, responses were evaluated on the basis of the above four reasoning levels according to standardized procedures.

Two scores were calculated. A map reading score (*MS*) was computed for each subject based on the number of items answered 'correctly,' i.e. in a manner similar to traditional scoring methods. All formal responses were counted as correct, as were some 11 concrete responses which would be acceptable answers on a regular geography test. To obtain a reasoning level (*RL*) measure the item distractors were assigned values of one through four in ascending order of development, and a reasoning level mean was computed for each subject with the following formula:

$$RL = \frac{(\text{Imaginary/Tautological} \times 1) + (\text{Perceptual/Associative} \times 2) + (\text{Concrete} \times 3) + (\text{Formal} \times 4)}{25 (\text{Number of Items})}$$

RL was compared with *MS* for reliability and relationship to other variables.

Hypotheses

Hypothesis 1 predicted that *RL* increases as grade level increases. This hypothesis was intended to test the validity of *RL* as a measure of cognitive development level. In previous studies (Feldman, 1969, 1970), *MS* was found to increase significantly with grade level; because *MS*, i.e. achievement, was supposed to be dependent upon increased cognitive development, *RL* should also increase with age.

Hypothesis 2 followed from the work of Turiel (1966, 1969) and others on stages of moral reasoning development. Turiel and his colleagues have found that responses to moral questions tend to be distributed in an almost gaussian fashion, with most responses being at a single stage, fewer responses at stages ± 1 stage from the modal level, still fewer responses more than ± 1 stage from the modal level (see Figure 1). It was predicted that *RL* responses would also tend to exhibit a modal dominant response stage, fewer responses more than ± 1 stages from the modal stage. Positive results for Hypothesis 2 would support *RL* as a valid measure of cognitive development, since Piaget (1950) argues for the underlying unity of all cognitive development. Thus, findings for spatial reasoning and moral reasoning should be related to general cognitive development stages in much the same way.

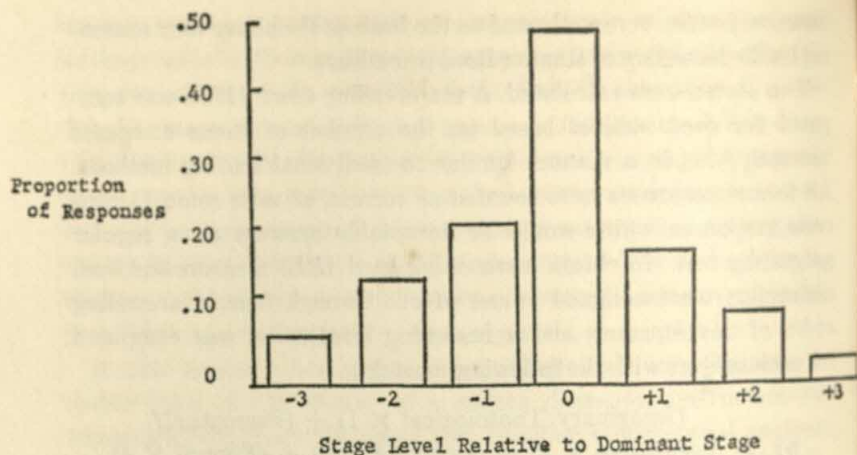


Figure 1. Profile of moral stage usage on Kohlberg moral judgment interview (from Rest, Turiel, and Kohlberg study; taken from Turiel, 1969).

Hypothesis 3 predicted that correlations of *MS* and *RL* with variables such as IQ, map drawing, sex, and ethnicity would not differ significantly. To an undetermined extent, *RL* and *MS* were artifactually related since they were computed from the same data. *MS* and *RL* for a given individual would have to be similar since those responses counted as *correct* were either formal or concrete; the very responses which would also contribute to a high *RL*. Still, *MS* and *RL* could and did vary in individual instances, as in the following hypothetical example:

S1: <i>MS</i> = 15	15 Formal × 4 points	= 60
	10 Concrete × 3 points	= 30
	0 Perceptual × 2 points	= 0
	0 Imaginary × 1 point	= 0
	<hr/>	
	<i>RL</i> = 3.80	= 90/25
S2: <i>MS</i> = 15	4 Formal × 4 points	= 16
	11 Concrete × 3 points	= 33
	0 Perceptual × 2 points	= 0
	10 Imaginary × 1 point	= 10
	<hr/>	
	<i>RL</i> = 2.36	= 59/25

Hypothesis 4 predicted that despite possible differences in achievement and reasoning levels according to ethnic background, all

groups could be shown to be proceeding toward formal thought through the same set of stages. Previous studies (Feldman, 1969, 1970) had found that when all 25 items of the map test were subjected to scalogram analyses, each of three ethnic groups' performance tended to form a scalable item set, but the items were acquired in a somewhat different order within each group. Previous research on sequences of cognitive development has supported Piaget's claim that stages of cognitive development are invariant (Kohlberg, 1968; Wallach, 1963). Feldman (1970), in reviewing the literature and in trying to explain his disparate results, concluded that individual differences are likely to affect stages of acquisition of reasoning skills when there are many specific tasks to test a limited span of development. Therefore, hypothesis 4 predicted that a reanalysis of Feldman's (1970) 25 step sequence in terms of Piaget's major stages would yield a more invariant sequence.

Method

Description of Subjects and Sample

Two samples, chosen for different purposes, were the sources of data for the study. A sample consisting of 270 fifth, seventh and ninth grade public school students of equivalent social class (lower working class) but differing in ethnicity was drawn to test the effects of ethnic differences and grade level on reasoning levels and sequences. Subjects were distributed evenly across three ethnic groups (Black, White, and Chinese) and the three grade levels. This sample was drawn from four public schools in San Francisco, California (for more details of the sample and sampling techniques, see Feldman, 1969, 1970). To gather data on the stability of the instruments, fourth and sixth grade students ($N = 88$) attending a St. Paul, Minnesota elementary school were sampled. This sample was chosen for its wide SES range and middle class bias, i.e., because it was more representative of typical school populations than the San Francisco sample.

Procedures for Administration

A primary purpose of the testing procedure was to reduce, within practical constraints, the dependence of a child's performance upon

his ability to read. Another purpose was to reduce the anxiety of test taking to a minimum so that each child had the best opportunity to exhibit his reasoning about spatial concepts.

All Ss were tested in the classroom by the same examiner (*E*) accompanied by one or two assistants. For most groups, *E* was of the same ethnic background as the Ss. After a brief informal warm up to establish rapport, *E* read all directions, each question and its distractors aloud calling attention to the option of *S* to write in his own answer, proceeding as slowly as was necessary to insure each child the opportunity to think about and complete the items. The *E* and assistant(s) answered all questions by Ss on a one-to-one basis when Ss indicated that they did not understand an item. Students were given individual assistance in expressing their own responses, and some who were obviously handicapped by inability to write with facility dictated oral responses to *E* or an assistant who then wrote down the response. All groups were told that they were "testing the test"—this idea seemed to appeal to them. On the whole Ss seemed to find the items interesting and appeared to be motivated to select the responses which seemed "best to them."

Instrument Validity and Reliability

The validity of the instrument was tested against a map drawing criterion in the San Francisco sample. The correlation between map drawing skill and map reading scores was found to be .46 ($p < .01$) for the entire sample; sub-group correlations did not significantly differ from the sample. Thus, it may be said that the instrument is a reasonably good predictor of a map-drawing criterion. The restricted variability in map categories (six categories) may have affected the map category \times map reading score correlation (see Feldman, 1969, 1970 for more details).

In the St. Paul sample, test-retest reliability coefficients of .74 for map reading scores (*MS*) and .81 for reasoning level (*RL*) were obtained. Practice effects appeared to be insignificant; a mean map score of 16.85 for two classes which had not taken the test in October ($N = 48$) compares closely with the retest map score mean of 17.03. Thus, the increase from 15.52 in map score mean in October to 17.03 for the retest in January can probably be attributed to causes other than practice effects (e.g., motivation, school experience).

TABLE 1

Map Score Means and Standard Deviations for the San Francisco Sample
($N = 270$)

Ethnic Group		Grade Level			Total
		5th	7th	9th	
Black	\bar{X}	11.97	13.40	15.17	13.51
	SD	(3.81)	(3.37)	(3.51)	(3.57)
White	\bar{X}	14.20	16.17	17.20	15.85
	SD	(3.57)	(3.15)	(3.67)	(3.47)
Chinese	\bar{X}	14.57	17.40	17.63	16.53
	SD	(3.07)	(3.43)	(3.39)	(3.30)
Total	\bar{X}	13.53	15.63	16.63	15.30
	SD	(3.50)	(3.32)	(3.52)	(3.86)

Results

Hypothesis 1

Hypothesis 1 predicted that RL would increase significantly with grade level. Table 1 shows MS means and SD s for the San Francisco sample ($N = 270$). Table 2 shows RL means and SD s for the same sample. Previous results (Feldman, 1970) have shown grade level to be a significant ($p < .01$) influence on MS ; Table 3 presents an analysis of variance testing the effects of grade level on RL . As can be seen from Table 3, grade level significantly influenced RL ($p < .01$), supporting Hypothesis 1.

Hypothesis 2

Hypothesis 2 predicted that S s would respond most frequently to distractors at the reasoning level hypothesized to be the S 's

TABLE 2

Reasoning Level Means and Standard Deviations for the San Francisco Sample
($N = 270$)

Ethnic Group		Grade Level			Total
		5th	7th	9th	
Black	\bar{X}	2.65	2.67	2.94	2.75
	SD	(0.33)	(0.36)	(0.34)	
Chinese	\bar{X}	2.93	3.14	3.13	3.07
	SD	(0.29)	(0.33)	(0.36)	
White	\bar{X}	2.85	2.95	3.16	2.99
	SD	(0.32)	(0.35)	(0.35)	
Total		2.81	2.92	3.07	2.94

TABLE 3

Analysis of Variance Testing Effects of Grade Level and Ethnic Group on Reasoning Level Means (N = 270)

Source	df	MS	F	p
Grade Level	2	2.37	14.81	<.01
Ethnic Group	2	1.63	21.55	<.01
GL \times EG	4	0.16	1.45	NS
Error	261	0.11		

dominant stage, less frequently ± 1 reasoning level away from the dominant stage, still less frequently more than ± 1 category away.

Figure 2 presents an analysis of cognitive stages which parallels Turiel's method of determining the profile of responses relative to a dominant stage. Ss were categorized into dominant cognitive stages on the basis of the following decision rule: Ss who had *RLs* of 1.00 to 1.75 were hypothesized as tending to respond at a tautological level; *RLs* of 1.76 to 2.50 at a perceptual level; *RLs* of 2.51 to 3.00 at a concrete level; *RLs* of 3.01 to 4.00 were considered to be responding formally (see Figure 3). These response ranges were selected somewhat intuitively; decisions were made on the basis of

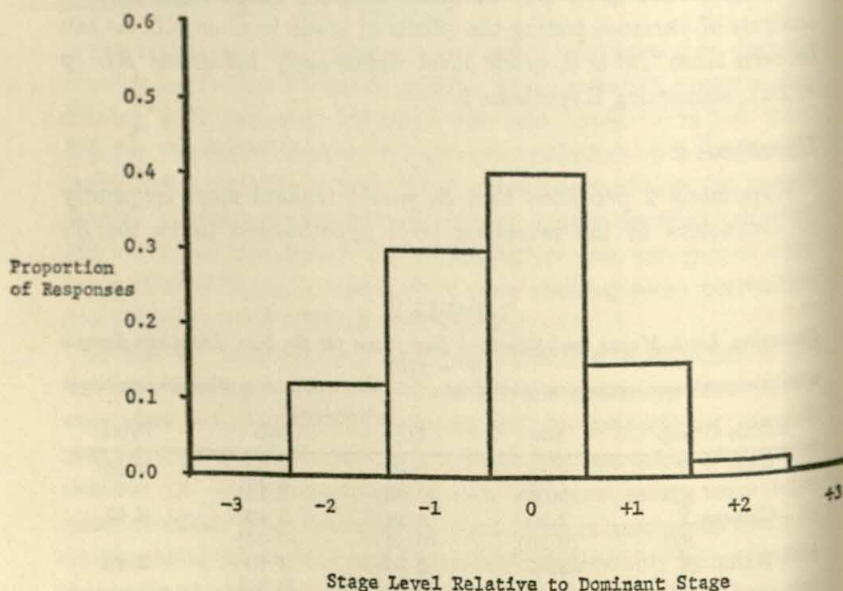


Figure 2. Profile of cognitive stage responses on Salomon/Feldman (1969) map reading test.

previous findings and best guesses (Feldman, 1969, 1970). On the basis of the categorizing method described above, no subjects fell into the Tautological stage, 33 fell into the Perceptual stage, 109 into the Concrete stage, and 128 into the Formal reasoning category. Although the resulting distribution of responses is flatter than that found by Rest, Turiel, and Kohlberg (1969), and although a restricted number of categories (4 versus 6) may have distorted the distribution somewhat, the data tend to support the prediction; i.e., responses did tend to fall into the hypothesized dominant response category. Obviously, one could adjust the *RL* ranges to produce increasingly better fits to the predicted distribution; however, only one algorithm was used in this study.

It should be noted that only at the Formal level is there necessarily an artifactual relationship between *RL* and the modal response to distractor types. As was illustrated above in two hypothetical cases, a given *RL* could be achieved with a variety of patterns of responses. Thus, it appears as if *RL* categorizing produced results consistent with those of previous research (Turiel, 1969; Rest, Turiel and Kohlberg, 1969), and also consistent with the cognitive-developmental theory of Piaget.

Hypothesis 3

Table 4 presents a correlation matrix of *MS*, *RL* and five other variables. Table 4 shows that *MS* and *RL* correlated .93. It was predicted that *MS* and *RL* would correlate with sex, IQ and grade level to about the same extent. In view of the high correlation between *MS* and *RL*, it was almost inevitable that the two measures would exhibit a similar pattern of intercorrelations with other variables. Thus, although Hypothesis 3 was supported, no clear indication of the extent to which map achievement (*MS*) and reasoning level are related to other variables will be available until *MS* and *RL* are assessed with different instruments. The only indication from previous research of the relationship between map reading and reasoning level (Salomon, 1968) is a .53 ($p < .01$) correlation between a variation of the map test used in the present study and a Piagetian spatial reasoning task (objects-on-a-slope). As reported above, *RL* was slightly more reliable over a three month interval than *MS* (.81 vs. .74).

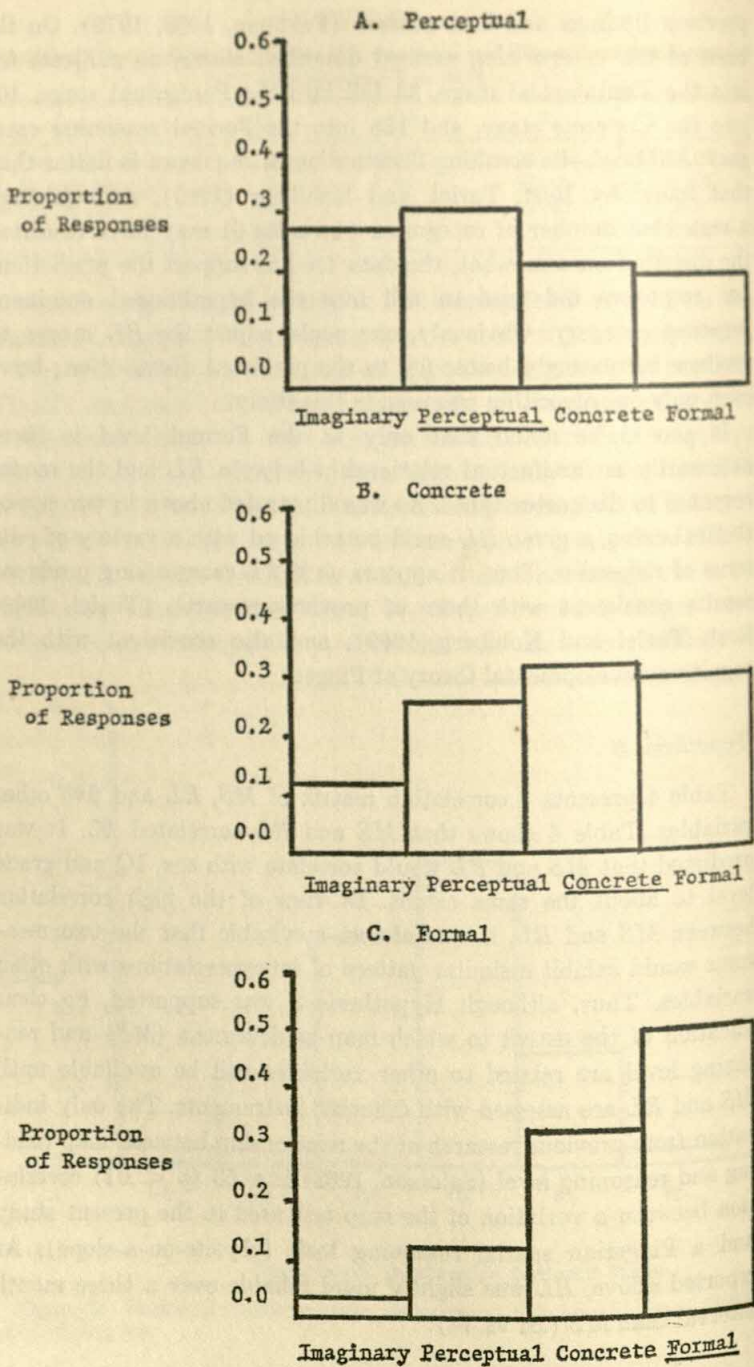


Figure 3. Distribution of responses to map test questions for subjects

TABLE 4

*Correlation Matrix of Map Score and Reasoning Level with Other Variables
(San Francisco Sample, N = 270*)*

Variable	Grade Level	Sex	Ethnic Group	IQ	Reasoning Level	Map Score	Map Category
Grade Level	—	.05	.00	.23*	.41**	.41**	.25**
Sex	—	—	-.07	.07	.02	-.01	-.19*
Ethnic Group	—	—	—	.22*	.21**	.22**	.22**
IQ	—	—	—	—	.58**	.60**	.40**
Reasoning Level	—	—	—	—	—	.93**	.48**
Map Score	—	—	—	—	—	—	.46**
Map Category	—	—	—	—	—	—	—

* Except for correlations involving IQ, where N = 199.

* p < .05.

** p < .01.

Hypothesis 4

Hypothesis 4 predicted that despite differences among ethnic groups in *MS* and *RL* (both significant), all groups go through the same set of developmental stages. Figure 4 shows the number of subjects at each reasoning level, with the sample analyzed first by grade level, then by ethnic group. As seen in Figure 4, 5th grade children tend to respond at a concrete level of reasoning, with smaller numbers responding predominantly perceptually and formally. At 7th grade, the number of perceptual respondents decreased slightly, the number of concrete respondents also decreased slightly, while formal respondents increased from 31 to 41 (out of 90 possible). At 9th grade, the number of *Ss* responding at perceptual and concrete levels continued to decrease, while 56 of 90 *Ss* tended to respond at formal reasoning levels. Thus, the predicted developmental changes with increased age were supported by the data, at least for the sample taken as a whole.

When the data were analyzed separately for each ethnic group, Black *Ss* (from all three grades) were distributed across the three reasoning stages in almost precisely the same numbers as were 5th grade *Ss* (see Figure 4). White *Ss* were distributed across reasoning stages in numbers similar to 7th grade children, and Chinese children were distributed across reasoning stages in numbers almost identical to 9th grade children. (It should be noted that the same sample was analyzed both for grade level and ethnic group differences; thus one third of the same subjects were in both sets of frequencies).

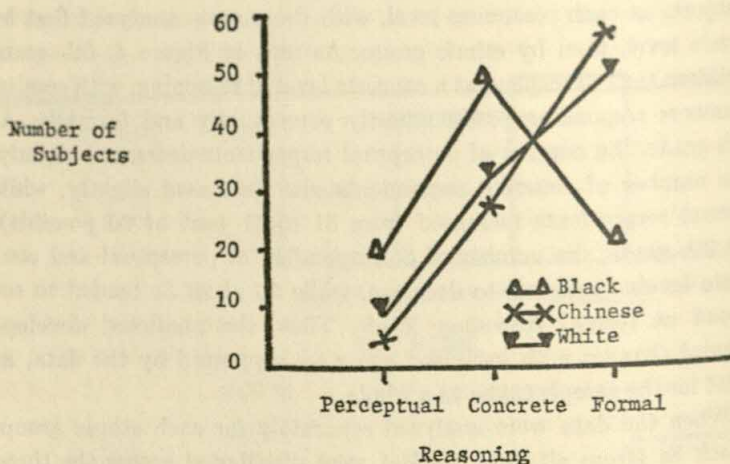
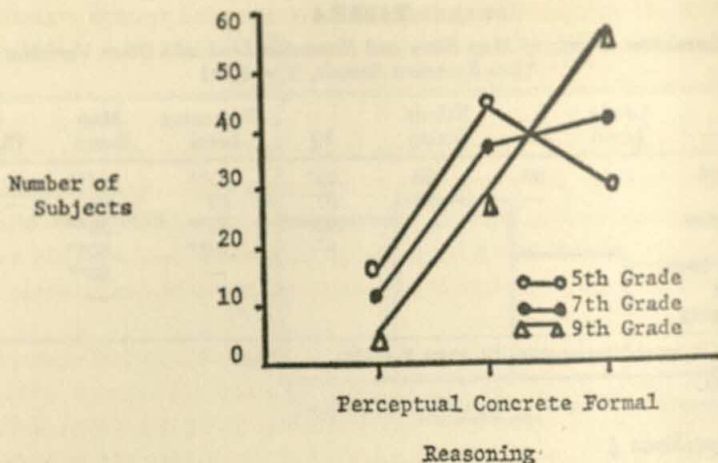


Figure 4. Number of subjects at each reasoning level stage analyzed by grade level and by ethnic group ($N = 270$).

From the similarity of distributions for each of the three ethnic groups to each of the three grade levels, it would appear that Hypothesis 4 was supported. It appears that the differences between ethnic groups are not due to fundamental differences in cognitive developmental stages themselves, but rather due to the age by

which the members of each group have achieved certain reasoning skills.

Discussion

The purpose of this study was to test the efficacy of using ranked distractors for the assessment of Piagetian reasoning levels. Since the data tended to support the hypotheses, and since the stability of the instrument was relatively high, it would appear that the map instrument may be capable of measuring reasoning stage levels as well as map achievement.

If, as the results indicate, levels of reasoning do increase with grade level, and if the rate of stage to stage development varies, the diagnosis of individuals' developmental levels using ranked distractors would appear to have potential importance for educators (Adler, 1963; Stone, 1966; Sullivan, 1967; Feldman, 1969, 1970). While a great deal of work has been done attempting to replicate Piaget's results, particularly in areas such as conservation and object permanence, little seems to have been accomplished in setting up standardized batteries for diagnostic purposes. The level of a student's or group's cognitive functioning is obviously of importance to the teacher—in planning curricula, instructional strategies, manipulating the educational environment, etc. Piaget's clinical methods are impractical because they require long periods of concentrated observation by trained observers, coupled with the use of non-standardized questioning (Ginsburg and Oppen, 1969). As a pencil-and-paper measure of reasoning level, the Feldman-Salomon technique of using ranked distractors has the advantage that it may be utilized in ordinary classroom situations; it requires a reasonable amount of time (60 to 90 minutes) and does not require a highly trained examiner.

Despite this seeming potential significance as a diagnostic instrument applied to the classroom situation, it is as a heuristic device that this measuring technique seems to offer the most promise at this time. Piaget has provided a wealth of material for hypothesis construction and experimentation (Flavell, 1963), but there is also, as Ausubel (1967) pointed out, some confusion among educators and students who are "bewildered by the overstated claims of Piaget's supporters and detractors regarding the applicability of his ideas to educational theory and practice." Thus, while the

exploration of practical applications of research on sequences of stage development theories should continue, there is a danger that these applications may be premature extrapolations (Sullivan, 1967, p. 17):

One note of caution should be sounded in the treatment of Piaget's observations as a structural theory of intelligence. Piaget's theoretical model of intellectual development is a rather elaborate superstructure which is superimposed on his observations. More research is needed to clarify how his 'structures' account for the observations that he has catalogued.

Thus, our research will continue to seek data bearing on the following questions:

1. What mechanisms affect cognitive development?
2. How do these mechanisms operate?

It would seem as if these questions must be understood if educators are to optimize progress through stage development. Four possible mechanisms have been suggested (Piaget, 1964, Sullivan 1967) which, singly or combined, could account for Piaget's observations on stage development; they are:

1. maturation,
2. interaction with the physical environment,
3. social interaction with peers,
4. equilibration.

The results reported here suggest that an examination of the types of responses chosen or volunteered by children at differing levels of development may provide insights into the operation of general mechanisms of cognitive development. Turiel (1966, 1969) and his colleagues have begun to explore this approach in their studies of moral development. Item response analysis takes on a new dimension in this context, since each response potentially reflects a specific level of reasoning. A battery of such instruments could be useful in establishing conditions for influencing interstage transitions, making comparative analyses of before-and-after treatment effects, and in providing data necessary for operationally defining development of cognitive processes.

It should be noted that the notion of "stage" is itself a heuristic

device to describe qualitative differences in reasoning. Cognitive development almost certainly does not take place in disjunct stages. As data from Rest, Turiel and Kohlberg (1969) and the present study indicate, responses tend to cluster about a modal stage but do not all fall in that stage. Researchers and educators should be cautioned not to categorize an individual into a given stage and assume he can respond *only* at that level. In view of the complexity of piagetian theory, and in view of the difficulty in rendering cognitive development researchable, the results found in this study should be construed only as observed behavioral differences which may or may not reflect developmental processes.

REFERENCES

- Adler, M. *Some implications of the theories of Jean Piaget and J. S. Bruner for education*. Toronto: Board of Education for the city of Toronto. Research Service, 1963.
- Ausubel, D. Foreward. In Sullivan, E. V. *Piaget and the school curriculum: A critical appraisal*. Toronto: The Ontario Institute for Studies in Education, Bulletin No. 2., 1967.
- Eisner, E. W. A comparison of the developmental drawing characteristics of culturally advantaged and culturally disadvantaged children. Project No. 3086, Contract No. OE 6-10. 027, Stanford University, 1967.
- Feldman, D. H. A study of a fixed sequence of skill and concept acquisition requisite to performance of a common school task: Map drawing. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, February 1969.
- Feldman, D. H. The fixed-sequence hypothesis: Individual differences in the development of school related spatial reasoning. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, March 1970.
- Flavell, J. H. *The developmental psychology of Jean Piaget*. New York: D. Van Nostrand Co., 1963.
- Gagné, R. M. The acquisition of knowledge. *Psychological Review*, 1962, 69, 355-365.
- Ginsburg, H. and Oppen, S. *Piaget's theory of intellectual development: An introduction*. Englewood Cliffs, N. J.: Prentice-Hall, 1969.
- Guttman, L. and Schlesinger, I. M. Systematic construction of distractors for ability and achievement test items. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 569-580.
- Kohlberg, L. Early education: A cognitive-developmental view. *Child Development*, 1968, 39, 1013-1062.
- Piaget, J. *The psychology of intelligence*. New York: Harcourt, Brace and World, 1950.

- Piaget, J. Cognitive development in children: The Piaget papers. In Ripple, R. E. and Rockcastle, V. N. (Eds.), *Piaget rediscovered: A report of the conferences on cognitive studies and curriculum development*. Ithaca, New York: School of Education, Cornell University, 1964, pp. 6-48.
- Piaget, J. and Inhelder, B. *The child's conception of space*. New York: W. W. Norton Co., 1967 (First published in 1948).
- Rest, J., Turiel, E., and Kohlberg, L. Level of moral development as a determinant of preference and comprehension of moral judgments made by others. *Journal of Personality*, 1969, 37, 2, 225-252.
- Salomon, G. Cultural differences in reading and understanding geographic maps. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1968.
- Salomon, G. and Feldman, D. H. *A map drawing and map reading exercise*. Stanford, California: Stanford Center for Research and Development in Teaching, Copyright 1969.
- Stone, M. *The Development of the intersituational generality of formal thought*. Unpublished doctoral dissertation, University of Illinois, 1966.
- Sullivan, E. V. *Piaget and the school curriculum: A critical appraisal*. Toronto: The Ontario Institute for Studies in Education, Bulletin No. 2. 1967.
- Turiel, E. An experimental test of the sequentiality of developmental stages in the child's moral judgments. *Journal of Personality and Social Psychology*, 1966, 3, 611-618.
- Turiel, E. Developmental processes in the child's moral thinking. In P. H. Mussen, J. Langer and M. Covington (Eds.), *Trends and Issues in Developmental Psychology*, New York: Holt, Rinehart & Winston, 1969, 92-131.
- Wallach, M. A. Research on children's thinking. In H. Stevenson (Ed.), *Child Psychology*. Chicago: University of Chicago Press, 1963.

NOTES ON APPROXIMATE PROCRUSTES ROTATION TO PRIMARY PATTERN¹

ESKO KALIMO

The National Pensions Institute of Finland

AN increasingly popular trend in confirmatory factor analysis is rotating to a maximum fit to a specified factor matrix. This kind of rotation to a target matrix, instead of the more usual criterion to a maximum simple structure, has been called the "Procrustes" approach. The problem was first considered by Mosier (1939), who presented a mathematical rationale for oblique rotation to a target matrix, expressed as a reference structure. Because he found that the equations were difficult to solve algebraically, he suggested an approximate solution for rotating a factor matrix to the best least squares fit to the target reference structure. The same approximate solution was later suggested by Horst (1956) and Hurley and Cattell (1962). The authors of the last-mentioned article relied partly on the work of Ahmavaara (1954), who used the same formulas for a closely related transformation: for comparing two factor analyses with each other. Lubin (1950) proposed essentially the same method to be used with Eysenck's "criterion analysis" (1950). Harman (1967, p. 251) refers to this procedure as a general method giving the matrix of transformation between any two solutions in the same common factor space.

¹ This paper was written while the author was at the School of Hygiene and Public Health, The Johns Hopkins University. The work was supported by Grants (8 R01 HS 00110 and 8 T01 HS 00012) from the National Center for Health Services Research and Development, (5 D04 AH 00076) from the National Institutes of Health, U. S. Department of Health, Education and Welfare, and from the Commonwealth Fund of New York.

The author is indebted to Dr. Bert F. Green for his helpful comments on an earlier draft of the manuscript.

This approximate solution for a maximum fit to a target matrix has been applied to target matrices, expressed as reference structures. Since it is often conceptually and computationally simpler to use primary patterns, the procedures dealing with reference structure have made additional computations necessary to get the final matrix as a primary pattern (see e.g., Hendrickson and White, 1964). A direct procedure to a primary pattern, when the target matrix is a reference structure, will be suggested. Secondly, it will be shown that the rotated primary patterns are not usually identical when the target matrix is a reference structure and when it is the corresponding primary pattern.

It may be noted that, under the restriction of orthogonal rotation, exact solutions for Procrustes rotation have been given by Green (1952), Cliff (1966), and Schönemann (1966). Browne (1967) suggested an exact solution for the Procrustes rotation to oblique reference structure. Fischer and Roppert (1964), Browne and Kristof (1969), and Gruvaeus (1970) have developed exact Procrustes methods to oblique primary pattern, by minimizing slightly different criteria than Mosier (1939). In spite of these recent developments, it seems to be well-founded to examine further the properties of the approximate oblique procedure, especially because of its simplicity. Neither can this approximate method be abandoned before more research has been done on the superiority of the exact methods in practice.

Procedures

Consider an unrotated orthogonal factor matrix for n variables and m factors, A ($n \times m$), which we would like to rotate to a maximum fit to a target matrix, B (usually $n \times m$). Let L ($m \times m$) stand for the desired transformation matrix and E ($n \times m$) for the matrix of differences between the rotated matrix and the target matrix, represented by

$$E = AL - B. \quad (1)$$

The sum of squares, $tr(EE')$, is an appropriate criterion to be minimized. As Mosier showed, the transformation matrix which gives a least squares fit of matrix A to matrix B is given by

$$L = (A'A)^{-1}A'B. \quad (2)$$

When the target matrix is a reference structure, V ($n \times m$), and if

we require that the factors have unit variances, matrix L must be normalized by columns to give the final transformation matrix, Λ ($m \times m$), which would satisfy the restriction

$$\text{diag} (\Lambda' \Lambda) = I. \quad (3)$$

Then

$$A \Lambda = V. \quad (4)$$

By standard techniques (Harman, 1967), matrix V can be transformed to the corresponding primary pattern and the correlations among the primary factors calculated. The well-known relation between the unrotated factor matrix and a primary pattern, P ($n \times m$), is expressed by

$$A T^{-1} = P, \quad (5)$$

in which T^{-1} ($m \times m$) refers to the transformation matrix. Its inverse T ($m \times m$) gives the direction cosines of the primary factors with respect to the unrotated factors and is subject to the following restriction, in the usual case of factors with unit variances,

$$\text{diag} (T T') = I. \quad (6)$$

The well-known relation between matrices T and Λ is

$$T = D_1 \Lambda^{-1}, \quad (7)$$

in which D_1 ($m \times m$) refers to a diagonal matrix with elements for normalizing matrix Λ^{-1} by rows. These elements are also the correlations between the corresponding rotated primary and reference factors. When equation (7) is expressed in terms of matrix L , we get

$$T = D_1 (L D_2)^{-1}, \quad (8)$$

in which D_2 ($m \times m$) refers to a diagonal matrix with elements for normalizing matrix L by columns. This can be written

$$T = D_1 D_2^{-1} L^{-1}. \quad (9)$$

Because the elements of matrix D_1 are always chosen so that matrix $D_2^{-1} L^{-1}$ becomes normalized by rows, matrix T can be directly calculated by

$$T = D_3 L^{-1}, \quad (10)$$

in which D_3 ($m \times m$) refers to a diagonal matrix with elements for

normalizing matrix L^{-1} by rows. The final transformation matrix, T^{-1} , is given by

$$T^{-1} = LD_3^{-1}. \quad (11)$$

This formula shows a direct procedure for calculating the best fitting primary pattern, when the target matrix is expressed as a reference structure. The matrix of transformation to the best fitting primary pattern can thus be obtained directly by another simple modification of matrix L as the matrix of transformation to the corresponding reference structure. It is not necessary to proceed through the reference structure to be able to express the results as a primary pattern, for this method gives a primary pattern which is identical to that which would be obtained through the reference structure. So, if A is first rotated to a reference structure, using (2) and (4), it can later be rotated to the corresponding primary pattern either by applying the Procrustes method again, i.e., formulas (11) and (5), or by a transformation of the rotated reference structure using the standard technique.

The above result may also be interpreted so that, when the approximate Procrustes procedure is applied with a reference structure as the target, both the rotated reference structure and the corresponding rotated primary pattern are in the same sense maximally similar to the target matrix. It can be shown in a corresponding way that this is true also when we have a primary pattern as the target matrix, assuming that matrix L is again modified in the same ways, for producing the final transformation matrices which fulfill the restrictions imposed on them by the factor analysis model. However, when the exact Procrustes methods are used, the rotated reference structure and the corresponding primary pattern are not always in the same sense maximally similar to the target matrix (Browne and Kristof, 1969).

It can be shown, however, that when applying the approximate Procrustes method, the rotated primary pattern which is maximally similar to a target reference structure is not identical with the rotated primary pattern which is maximally similar to the corresponding target primary pattern.

When the target matrix is a reference structure, the matrix of transformation to a primary pattern can be written according to (9) and (11).

$$T^{-1} = LD_2D_1^{-1}. \quad (12)$$

The well-known relation between a reference structure and the corresponding primary pattern is

$$V = PD_4, \quad (13)$$

in which D_4 refers to a diagonal matrix giving the correlations between the corresponding reference and primary factors. According to (2) and (13), (12) can be written

$$T^{-1} = (A'A)^{-1}A'PD_4D_2D_1^{-1}. \quad (14)$$

It may be noted that D_4 refers to the correlations between the target reference and primary factors, while D_1 gives the correlations between the rotated reference and primary factors.

When the target matrix is a primary pattern, the matrix of transformation to the rotated primary pattern is according to (11)

$$T_p^{-1} = L_pD_3^{-1}. \quad (15)$$

D_3 in (15) is not usually equal to $D_1D_2^{-1}$ in (12) and (14), because they refer to normalizations of different matrix L^{-1} . By using (2) we get

$$T_p^{-1} = (A'A)^{-1}A'PD_3^{-1}. \quad (16)$$

According to formulas (14) and (16), matrices T_v^{-1} and T_p^{-1} are not usually identical. However, since the difference in their formation can be expressed as a diagonal matrix, the corresponding columns of T_v^{-1} and T_p^{-1} are proportional to each other. The resulting primary patterns are thus different depending on the form of the target matrix, but also their corresponding columns are proportional to each other. These pattern matrices differ generally only little from each other, leading in most cases to the same substantive conclusions.

When the target matrix is an orthogonal matrix and if we do not impose the restriction to orthogonality on the rotated factors, the Procrustes rotation can be carried out by the formulas given above either to a reference structure or to the corresponding primary pattern, depending on the preference of the factor analyst. Applications of the approximate Procrustes rotation with orthogonal target matrices are published elsewhere (Bice and Kalimo, in press).

Summary

After a brief review of rotation methods to a specified target matrix in factor analysis, two notes were presented. The notes dealt with the approximate Procrustes method, which rotates a factor matrix to an approximate least squares fit to a target matrix. The first note suggested a direct procedure for obtaining a primary pattern, when the target matrix is a reference structure. On the basis of this result it was also concluded that the resulting primary pattern and the corresponding reference structure are in the same sense maximally similar to the target matrix. The second note showed that usually numerically different but interpretatively similar primary patterns are obtained when the target matrix is a reference structure and when it is the corresponding primary pattern.

REFERENCES

- Ahmavaara, Y. *Transformation analysis of factorial data*. Helsinki: Suomalainen Tiedeakatemia, 1954.
- Bice, T. and Kalimo, E. Comparisons of health-related attitudes: A cross-national, factor-analytic study. *Social Science & Medicine*, in press.
- Browne, M. W. On oblique Procrustes rotation. *Psychometrika*, 1967, 32, 125-132.
- Browne, M. W. and Kristof, W. On the oblique rotation of a factor matrix to a specified pattern. *Psychometrika*, 1969, 34, 237-248.
- Cliff, N. Orthogonal rotation to congruence. *Psychometrika*, 1966, 31, 33-42.
- Eysenck, H. J. Criterion analysis—An application of the hypothetico-deductive method to factor analysis. *Psychological Review*, 1950, 57, 38-53.
- Fischer, G. and Roppert, J. Bemerkungen zu einem verfahren der transformations-analyse. *Archiv für die Gesamte Psychologie*, 1964, 116, 98-100.
- Green, B. F. The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika*, 1952, 17, 429-440.
- Gruvaeus, G. T. A general approach to Procrustes pattern rotation. *Psychometrika*, 1970, 35, 493-505.
- Harman, H. *Modern factor analysis*, 2nd ed. Chicago: University of Chicago Press, 1967.
- Hendrickson, A. E. and White, P. O. Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, 1964, 17, 65-70.
- Horst, P. A simple method of rotating a centroid factor matrix to a simple structure hypothesis. *Journal of Experimental Educa-*

- Hurley, J. R. and Cattell, R. B. The Procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science*, 1962, 7, 258-262.
- Lubin, A. A note on "criterion analysis." *Psychological Review*, 1950, 57, 54-57.
- Mosier, C. I. Determining a simple structure when loadings for certain tests are known. *Psychometrika*, 1939, 4, 149-162.
- Schönemann, P. H. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 1966, 31, 1-10.

COMMUNALITY ESTIMATION IN FACTOR ANALYSIS OF SMALL MATRICES

EDWARD E. CURETON

University of Tennessee

NOWADAYS factor analyses of medium to large matrices are done on electronic digital computers. But small matrices can still be factored efficiently by the complete centroid method on desk calculators. It is not commonly noted, however, that the required accuracy in communality estimation *increases* as the number of variables *decreases*. Yet in centroid analysis, this fact is fairly obvious. In a small matrix, the error in the estimate of a diagonal entry is a substantial fraction of the column sum on which the factor loading is based. In a large matrix it is not.

In all too many factor analyses of small matrices, the investigators take as initial estimates of communality the absolute value of the numerically highest correlation in each column ($|r| \text{ max}$), and repeat this procedure with each residual matrix. In so doing they overlook or ignore Thurstone's explicit warning (1947, p. 300), "This simple method of estimating communalities is useful only for large correlation matrices. It is not applicable to small tables."

Of the more elaborate methods of communality estimation which are still within practical limits for desk calculators, the one most commonly recommended is the "miniature centroid" method (Medland, 1947; Thurstone, 1947, p. 300, eq. 15). Each variable is grouped with the two to four others with which its correlations are highest, the highest correlation in each column is placed on the diagonal, and the communality of the variable is taken as the square of its first centroid factor loading. With small matrices, this method also is unsatisfactory. It assumes that each miniature centroid matrix will have rank almost unity, but in small matrices it often happens that the two to

four variables with which a given variable correlates highest do not form with it a submatrix of rank close to unity.

If two variables have vectors which are close together in the commonfactor space, their correlation will be close to the geometric mean of their communalities. The correlation will underestimate one communality and overestimate the other. Cattell (1952, pp. 154-5) ascribes to Burt (1940) the suggestion that if the column sum for a given variable is high, the highest- r estimate of communality is likely to be too low, while if the column sum is low, the highest- r estimate is likely to be too high. Cattell noted also that the dispersion of the highest r 's is usually less than that of the final computed communalities. These observations are the basis of the method of initial communality estimation proposed here.

If each highest r is weighted by the corresponding column sum and divided by the sum of all column sums, the dispersion of the resulting estimates is about the same, on the average, as the dispersion of the final computed communalities. But in some individual cases the weighting correction goes in the wrong direction, and this weighting system gives *maximum* discrepancies (between estimated and final computed communalities) which in general are larger than those given by the unweighted highest- r method. We need to weight the highest r 's by quantities proportional to the column sums, but with less relative variability than that of the column sums themselves.

Two additional points need to be noted. First, if the correlation matrix contains negative entries, even if not enough to require reflection, the column sums needed for weighting are the sums of the absolute values of the correlations in the columns, not their algebraic sums. A variable with a low algebraic sum but a high absolute sum will have a low loading on the first factor, but high loadings on one or more later factors, and hence a high communality. Second, it is common lore that, on the average (over different factor analyses), the sum of the highest r 's gives a slight overestimate of the sum of the final communalities. Some empirical evidence, presented later herein, suggests that the total final communality is, on the average, about 96 per cent of the sum of the highest r 's. The proposed procedure is, then, as follows:

1. In the correlation matrix with empty diagonal, find the absolute column sums, $\Sigma |r|$. (Σ designates a sum of $n - 1$ quantities, where n is the number of variables.)

2. Add to each $\Sigma' |r|$ the *mean* of the n values of $\Sigma' |r|$, and call this sum S .

3. Record the absolute value of the highest r in each column, $|r|$ max.

4. Form the products $S|r|$ max.

5. The estimated communalities are then given by

$$C = S |r| \max \left(\frac{.96 \sum |r| \max}{\sum (S |r| \max)} \right). \quad (1)$$

Here Σ represents a sum of n quantities.

6. Estimate the reliabilities of the variables. If they are tests, and no better estimates are available, use Lord's (1959) empirical estimate of the standard error of measurement, from which

$$r_{11} = 1 - .187k/s^2, \quad (2)$$

where k is the number of items and s^2 is the variance of the scores. Then in any case in which the communality as estimated by (1) exceeds the estimate of reliability, substitute the latter.

Empirical Verification

In order to test the usefulness of (1), factor analyses of ten small matrices were studied.

1. Harman (1967, pp. 88-9) gives intercorrelations among six hypothetical variables which yield exact two-factor communalities.

2. Harman (1967, pp. 147, 154) also gives intercorrelations among eight physical measurements, and a principal-axes factor analysis to two factors, starting with computed communalities from the original study, which included 17 variables, by Mullen (1939). The largest discrepancy between any estimated communality and the final computed communality was .058.

3. Cureton et al. (1944) give intercorrelations among five verbal tests. The writer factored them by the centroid method to two factors, starting with communality estimates based on a tetrad-triad analysis. The largest discrepancy between any estimated communality and the final computed communality was .004.

4. Harman (1967, pp. 178, 186) gives intercorrelations among thirteen psychological tests, and a centroid factor analysis to three factors, starting with computed communalities from a previous bi-factor solution which included 24 tests. The largest discrepancy between any

estimated communality and the final computed communality was .045.

5. Lawley (in Thomson, 1951) gives intercorrelations among eight tests, and a maximum-likelihood factor analysis converging to exact sample communalities for two factors. The Lawley test did not reject the two-factor hypothesis at the .05 level, but did reject it at the .10 level. The writer, therefore, extracted a third factor by the centroid method, using column means after reflection (see later discussion) as residual communality estimates. This third factor was of doubtful significance, with highest loading .206 and total added communality .116, but it was retained for purposes of comparison of communality estimation methods, because with it the total communality was 94.8 per cent of the sum of the highest r 's, while without it the total communality was only 91.8 per cent of the sum of the highest r 's.

6. Lawley and Maxwell (1963, pp. 33-40) report intercorrelations among six sets of school grades, and a centroid factor analysis to two factors, repeated three additional times to improve the communality estimates. The Lawley test did not reject the two-factor hypothesis at the .50 level.

7. Lawley and Maxwell (1963, pp. 17-20, 24-27) report also intercorrelations among nine variables, and maximum-likelihood solutions for two and three factors. The Lawley test rejected the two-factor hypothesis at the .10 level but just failed to reject it at the .05 level. The three-factor hypothesis was not rejected at the .80 level. The three-factor solution was therefore used for our comparisons.

8. Swineford (1948) reports intercorrelations among nine tests, and gives a bi-factor solution. The writer factored this matrix to three factors by the centroid method, using Swineford's computed bi-factor communalities as initial estimates, and then repeated the centroid analysis six times. There was some doubt concerning the possible significance of a fourth factor, but a four-factor principal-axes solution, repeated once, was rejected because it gave one communality of .945 for a test whose reliability as estimated by Swineford was only .915.

9. Morrison (1967, pp. 243, 273-5), reports intercorrelations among the lengths of six chicken bones from a study by Wright (1954), and gives a maximum-likelihood factor analysis to two factors. The two-factor hypothesis was rejected by the Lawley test at the .001 level, so the writer extracted a third factor by the centroid method, using

column means after reflection as residual communality estimates. The highest third-factor loading was .206, all others were below .170, and the added total communality was .092, so it seemed clear that a fourth factor must be insignificant.

10. Fruchter (1954), pp. 72-85) reports intercorrelations among eleven Air Force tests, and a factor analysis by the centroid method, using Thurstone's procedure of estimating communalities by highest r 's and re-estimating by the same method in every residual matrix. Using several of the older approximate methods for determining the number of factors, he concluded that five were significant. Since he had only eleven variables, the writer suspected that he might have over-factored this matrix because of the highest- r re-estimates in the residual matrices. The writer therefore re-factored his matrix by the principal-axes method, taking as initial communality estimates values proportional to the squared multiple correlations but with sum equal to the sum of the highest r 's. A residual communality was replaced only if it was lower than one-half the mean absolute value of the off-diagonal entries in the column, in which case this latter value was used. It appeared that three factors were probably sufficient. The sums of squares of the factor loadings (eigenvalues for the principal axes solution) were as follows:

Factor	1	2	3	4	5	6	7
Eigenvalue	3.039	1.328	.666	.180	.132		
$\sum f^2$ centroid	3.018	1.293	.741	.248	.220	.096	.074

For the principal axes solution, the fourth eigenvalue is lower than the fifth Σf^2 of the centroid solution, and the fourth principal-axes factor had only two loadings above .20, neither of which was as high as .22. With three principal-axes factors, moreover, the largest discrepancy between an estimated communality and the final computed communality was less than .03. Communalities based on these three factors were therefore used in the comparisons.

Due to very large $N(8, 158)$, more than three factors would certainly have been statistically significant in Fruchter's data, and we may well have been unduly conservative in stopping at three. On the other hand, a factor of quite doubtful significance was added to the chicken bone data and another not much more significant to the Lawley-Thomson data, so perhaps for the purpose of communality

comparison we are justified in being conservative as to the number of factors in one of the ten samples.

The comparisons are shown in Table 1. The first three lines show the number of subjects N , the number of variables n , and the number of factors m , for each of the ten studies. We designate a final computed communality as h^2 , an estimated communality by (1) as C , and a highest- r estimated communality as r .

The fourth row of Table 1 shows $\Sigma h^2 / \Sigma r$ for each sample. The mean of the ten values is .962, and the standard error of this mean is .007. We conclude that for small matrices, the total communality is roughly 96 per cent of the sum of the highest r 's; hence the factor .96 in (1).

The next three rows show the within-sample ranges of the h^2 's, the C 's, and the r 's. For all samples but one, the range of the h^2 's is greatest, the range of the C 's is next, and the range of the r 's is least. In the one exception, underlined, the range of the C 's is greater than the range of the h^2 's.

The next two rows show the sums of absolute values of the discrepancies between h^2 and C , and between h^2 and r . In all cases except the one underlined, the sum of absolute values of the discrepancies is smaller for the C 's than for the r 's. By the one-sided sign test, this result is significant at almost exactly .01.

The last two rows show for each sample the one largest discrepancy between an h^2 and a C , and the one largest discrepancy between an h^2 and an r . In seven of the ten samples, the largest C -discrepancy is smaller than the largest r -discrepancy. In the other three cases, the C -discrepancy is underlined. While this result is far from conclusive for ten samples, it is at least indicative (significant at the .17 level by the one-sided sign test).

We conclude that the Burt-Cattell recommendation has been reduced to a fairly serviceable fixed procedure, and that (1) is to be preferred to the highest- r method for initial communality estimation. As compared to the latter, the use of (1) reduces the mean absolute value of the discrepancies between estimated and final computed communalities, and does not in general increase the size of the largest discrepancy.

Re-estimation

It is highly probable that Thurstone's caution was directed pri-

TABLE 1
Comparisons of Two Methods of Communalities Estimation

	Harman Artificial	Eight Phys. Meas.	Five Verbal Tests	Thirteen Psychol. Tests	Lawley- Thomson	Lawley- Maxwell Centroid	Lawley- Maxwell M-L	Swine- ford Centroid	Wright Chicken Bone	Fruchter Air Force
N	—	305	841	145	443	211	220	504	276	8,158
n	6	8	5	13	8	6	9	9	6	11
m	2	2	2	3	3	2	3	3	3	3
$\Sigma h^2 / \Sigma r$.960	.938	.984	.944	.948	.954	.957	.973	1.014	.948
Range h^2	.730	.371	.161	.523	.469	.243	.440	.354	.455	.364
Range e	.590	.309	.067	.522	.302	.255	.417	.321	.443	.318
Range r	.500	.252	.039	.405	.214	.185	.321	.260	.356	.310
$\Sigma h^2 - C $.344	.425	.182	.900	.566	.181	.544	.485	.612	.307
$\Sigma h^2 - r $.360	.639	.184	.742	.702	.230	.621	.547	.632	.486
Max $ h^2 - C $.118	.129	.058	.118	.224	.067	.152	.125	.439	.074
Max $ h^2 - r $.120	.177	.079	.150	.200	.091	.119	.107	.350	.105

marily toward re-estimation by highest r 's rather than to initial estimation by this method.

If we start a factor analysis of fallible data, knowing in advance the number of significant factors, and with communalities which are exactly correct for that number of factors (as, e.g., by re-factoring until the communalities are completely stabilized), and if we now factor with no revision of diagonal residuals, the following effects will be observed:

1. In each residual matrix, the diagonal elements will in general be smaller, relative to the off-diagonal elements, than in the previous matrix.

2. In the residual matrix after the last factor, the diagonal elements will all be exactly zero (within *rounding* error).

In each successive matrix, the real common variance is reduced, but the error variance is merely re-shuffled. The residual matrix after the last factor *should* have diagonal entries comparable in absolute value to the side entries if it is in fact an error matrix. About half of them should be positive and about half negative, with mean closer to zero than the individual entries. Re-factoring to complete stability represents over-fitting of the factors, with all error variance forced into the side entries in order to permit the final diagonal entries to be all exactly zero. This procedure tends to inflate the range of the communalities, leading occasionally to an artificial Heywood case, and even more often to a quasi-Heywood case, with one communality larger than the corresponding reliability.

In the correlation matrix, the initial communalities are comparable in magnitude to the highest r 's. In the last residual matrix from which a factor is extracted, they should be comparable in magnitude to the means of the entries in the columns, but must all remain positive. And in the residual matrix after the last factor, about half of them should be negative.

Since small matrices seldom have more than four or five factors, the following three rules will probably be sufficient:

1. Make initial communality estimates by (1).

2. For factors other than the first factor and the last factor retained, consider each diagonal residual, the absolute value of the highest residual correlation in each column, and each column mean, exclusive of the diagonal entry, after reflection. Choose the median of these three as the estimate of each residual communality.

3. For the last factor retained, estimate all residual communalities as the column means, exclusive of diagonal entries, after reflection.

Rule 2 says in effect to use the diagonal residual unless either the highest r is lower or the column mean is higher. If the highest r is lower, the original estimate by (1) was probably too high, and if the column mean is lower, it was probably too low.

Rule 3 recognizes the point that by the time the last factor to be retained is reached the highest r 's will be too high, and the diagonal residuals will consist mainly of the errors of estimation in the original communality estimates plus the errors of correction by Rule 2.

While these rules are rough, they do provide for relatively decreasing diagonal residuals in successive matrices, and permit some negative diagonal entries in the residual matrix after the last factor retained.

On the Number of Factors

The problem of the number of factors to be retained is usually less troublesome with small matrices than with larger ones, and the choice is commonly reduced to a choice between m and $m + 1$. No single rule is sufficient, but a judgment based on several will usually be correct.

1. A factor is likely to be significant if it has one loading as high as .30 and at least one other as high as .20, or if as many as one-fourth of its loadings are as high as .20, or if one-fifth are as high as .20 with at least one as high as .25.

2. A factor is likely not to be significant if the highest and the second highest loading of every variable are on preceding factors.

3. When the last factor is reached, the sum of squares of all loadings will usually reach 95 per cent of the initial trace (the sum of the estimated communalities). If this sum exceeds 105 per cent of the initial trace, the factor is probably not significant.

4. If N is at least 100 but less than 800, the Burt standard error of a factor loading (Burt and Banks, 1947) may be of some value. This test may be simplified by incorporating into it the hypothesis that the true loading is zero, yielding

$$\sigma_r = \sqrt{n/[N(n - m - 1)]}, \quad (3)$$

where the m -th factor is the one whose significance is to be estimated. The factor will usually be significant if more than one-fourth

its loadings exceed $2\sigma_f$, or if at least one-fifth exceed $2\sigma_f$ with one larger than $3\sigma_f$. If N is less than 100, a factor may be significant even when the Burt test says it is not, and when N exceeds 800, a factor may not be significant even when the Burt test says it is. In using the Burt test, estimate all diagonal residuals by Rule 3. The Burt test is about as good as any of the other approximate tests, and is easier to apply than most of them.

These tests are all, or almost all, inconsistent with one another. Nevertheless, if all of them are considered, the investigator can usually arrive at a correct conclusion.

REFERENCES

- Burt, C. *Factors of the mind*. London, University of London Press, 1940.
- Burt, C. and Banks, C. A factor analysis of body measurements for British adult males. *Annals of Eugenics*, 1947, 13, 238-256.
- Cattell, R. B. *Factor analysis*. New York, Harper & Brothers, 1952.
- Cureton, E. E., Cummins, Edith, Cynamon, M., Katzell, R. A., and Witmer, Louise R. Verbal abilities experiment: Analysis of new word meaning and verbal analogies tests. *P.R.S. Report No. 548*, Personnel Research Section, A.G.O. War Department, 1944 (mimeo).
- Fruchter, B. *Introduction of factor analysis*. New York: D. Van Nostrand Co., 1954.
- Harman, H. H. *Modern factor analysis*, Second Ed. Chicago: University of Chicago Press, 1967.
- Lawley, D. N. and Maxwell, A. E. *Factor analysis as a statistical method*. London: Butterworths, 1963.
- Lord, F. M. Tests of the same length do have the same standard error of measurement. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1959, 19, 233-239.
- Medland, F. F. An empirical comparison of methods of communality estimation. *Psychometrika*, 1947, 12, 101-109.
- Morrison, D. F. *Multivariate statistical methods*. New York: McGraw-Hill Book Co., 1967.
- Mullen, F. *Factors in the growth of girls seven to seventeen years of age*. Unpublished Ph.D. Thesis, Department of Education, University of Chicago, 1939.
- Swineford, F. *The nature of the general, verbal, and spatial bi-factors*. Supplementary Educational Monographs No. 67, University of Chicago Press, 1948 (special-group correlations, p. 14).
- Thomson, G. H. *The factorial analysis of human abilities*, Fifth Ed. New York: Houghton Mifflin Co., 1951.
- Thurstone, L. L. *Multiple factor analysis*. Chicago: University of Chicago Press, 1947.
- Wright, S. The interpretation of multivariate systems. In Kempthorne, O., et al., Eds. *Statistics and mathematics in biology*. Iowa: Iowa State University Press, 1954, pp. 11-33.

A HIGHER-ORDER ALPHA FACTOR ANALYSIS OF INTEREST, PERSONALITY, AND ABILITY VARIABLES, INCLUDING AN EVALUATION OF THE EFFECT OF SCALE INTERDEPENDENCY

RICHARD J. ROHLF¹

Guidance Bureau
Kansas University

REPORTED correlations among variables sampled from interest, personality, and ability measures, although often statistically significant, are usually too low to be of practical predictive value. The question of why the common variance among these three domains has been so low is intriguing. It was the thesis of this writer that the adoption of a hierarchical trait structure as a conceptual model would lend possible clarification to this question. Such a model postulates the existence of "higher-order" traits which account for, or explain, the covariation among traits of a "lower" level. Such higher-order constructs can conceivably have greater generalizability (Coan, 1964), and greater explanatory power (Royce, 1963) than the constructs at the next lower level. One purpose of this study was to explore the possibility of the existence of such a hierarchical trait structure through the use of a higher-order factor analytic procedure.

In this procedure the common variance among domains at the variable level can be accounted for in two ways. The first is through the composition of the first-order factors, e.g. individual first-order factors would be highly loaded on variables from separate domains.

¹ This article is based on a Ph.D. dissertation supervised by E. Gordon Collister. The author, now with the Counseling Center at Duke University, gratefully acknowledges the editorial assistance and support of Henry Weitz in the preparation of this article. All analyses were performed in the facilities of the Computation Center at the University of Kansas.

However, the findings of Bendig and Meyer (1963) suggest that interdomain relationships will not be accounted for in this way, if the degree of relationship among variables from different psychological domains is less than that which exists among variables from within the same domain. The findings of Bendig and Meyer, Anderson and Anker (1964), and Becker (1963) would suggest that this pattern exists among the intercorrelations of interest, personality, and ability variables.

Implicit in Bendig and Meyer's rationale is the notion that the interdomain variance at the variable level will instead be accounted for through the existence of correlated first-order factors. This is the second way in which the relationships between domains can be accounted for.

If the Bendig and Meyer thesis is correct (that a factor analysis of intercorrelations among variables from separate domains will yield correlated first-order factors, each of which being defined by a single domain), this would suggest a possible answer to the question posed above. The reason for the low intercorrelations among variables from these separate domains may be a result of dealing with constructs which are not of a "high" enough order. It would suggest the existence of constructs, specific to the three domains, that have greater generalizability and explanatory power than the constructs we are now measuring. If the correlations between such factors were of sufficient magnitude, it could possibly provide more satisfying evidence regarding hypothesized relationships among these three domains. Also, it would then be possible to obtain second-order factors which could be representative of a level of construct that would serve to further integrate the domains in question.

The Strong Vocational Interest Blank-Form M (SVIB) and the Minnesota Multiphasic Personality Inventory (MMPI) were selected to represent the areas of vocational interests and personality because of their extensive use in counseling and research. In selecting these instruments on this basis it was recognized that their use in a factor analysis was somewhat questionable. As Guilford (1952) pointed out, the interdependent nature of the scales in these two instruments results in partially "built-in" correlations among their respective scales. This built-in common variance among scales may, or may not, be an accurate reflection of the degree of relationship among

the variables in question for a given sample. Since the validity of this built-in variance must be held with some doubt for a given sample, any factor structure of these instruments must also be viewed with some degree of skepticism.

To what extent should one be concerned with this problem? This would be a function of the extent to which the observed correlations among the scales are a function of the empirical interdependence among the scales. If the built-in variance is extensive then one would, of course, be more concerned than if it is minimal. Therefore, a secondary purpose of this study was to determine the extent of the built-in correlations among the scales for both the MMPI and SVIB.

Havlicek (1965) and Shure and Rogers (1965) have pointed out that another consequence of these built-in relationships may be a distorted and an erroneous picture of factor stability. In order to investigate this question it was decided to determine the built-in factor structure of the MMPI and the SVIB and compare this structure with a structure based on a sample of subjects. If there were a good deal of similarity between the built-in factor structure and the sample structure, the hypothesis of an erroneous notion of factor stability would become more plausible. If there were little similarity between the two structures, it would lead to the inference that the independent portions of the scales are of greater influence in determining the factor structures of the instruments and the notion of a built-in factor stability would become less plausible.

Analysis of Sample Data

Subjects

The Ss of this study were those applicants for the Summerfield Scholarship at the University of Kansas who were chosen as finalists for the academic years, 1960-1961 ($N = 46$); 1961-1962 ($N = 47$); 1962-1963 ($N = 46$); 1963-1964 ($N = 40$); 1964-1965 ($N = 42$); 1965-1966 ($N = 34$); 1966-1967 ($N = 83$); 1967-1968 ($N = 62$). The Summerfield Scholarship is the highest academic honor that is awarded to a male undergraduate by the University of Kansas. These 400 candidates constituted the total number of Summerfield finalists, from the academic year of 1960-1961 through the academic year of 1967-1968, on whom the necessary data for this study were complete.

Instruments

The instruments which were used in this study were 45 scales of the SVIB-M, the three validity and 10 clinical scales of the MMPI, Terman's Concept Mastery Test (CMT), and the Stouffer Mathematics Test (SMT). This latter test is a locally devised instrument which is used in the selection procedures for determining the recipients of Summerfield Scholarship awards. The SVIB and MMPI data were in standard score form, and the CMT and SMT data were in raw score form.

Procedure

Prior to the factor analytic procedures, the distribution for each variable was tested for normality; a test for nonlinearity of regression was performed for every possible variable pair (Guilford, 1965, pp. 308-316), and Bartlett's test (Bartlett, 1950) was used to evaluate the significance of the study's original 60×60 intercorrelation matrix.

The factor analytic procedure for extracting higher-order factors consisted of the following:

- a. First, Kaiser and Caffrey's (1965) alpha factor analysis was performed on the original 60×60 intercorrelation matrix.
- b. This was followed by the orthogonal rotation of these alpha loadings to an approximate simple structure position using Kaiser's normal varimax criterion.
- c. These rotated varimax loadings were then further rotated to a maximum simple structure position using Eber's (1966) Maxplane oblique rotational solution.
- d. Three criteria, as suggested by Cattell (1966, pp. 188-189), were used to judge the adequacy of the final solution, including Bargmann's (1953) significance test for simple structure.
- e. The resulting correlations between factors were then evaluated to determine whether they were significantly different from zero, using Bartlett's test of significance.
- f. If the correlations proved to be significant, an alpha factor analysis was performed in order to extract second-order factors.

The same sequence of procedures, as listed above, was followed to complete the second-order analysis.

After the completion of the second-order analysis, a Cattell-

White transformation (Cattell, 1966, pp. 219) was performed to obtain the loadings of these higher-order factors on the original set of 60 variables.²

*Procedures for Determining the Extent of
Built-in Correlation Due to Item Overlap and K-Correction*

In order to assess the extent of the built-in correlation due to item overlap, 2500 randomly answered MMPI's and SVIB's were scored. The probability of occurrence of each of the three response alternatives (L, I, D,) on the SVIB was set at .333. The probability of occurrence of a true or false response on the MMPI was .5. Thus, each response alternative on the SVIB, and on the MMPI, had an equal probability of endorsement, i.e. any one of the thousands (3^{400} and 2^{566}) of unique response patterns possible on both instruments had an equal probability of occurrence as any other pattern of responses. The standard scores of these 2500 randomly answered inventories were then intercorrelated. The CMT and SMT were not included in this intercorrelation matrix because of their empirical independence of all the other scales.

It should be recognized that the method for determining the extent of built-in correlation among scales is representative of an overall "average" built-in effect. In order to calculate the built-in effect operating for a given sample, one would have to obtain the distribution of endorsements for each response alternative for each item. The probability of occurrence of each response alternative for a given item would then be based on the distribution of endorsements for that item rather than being set at an equal probability of occurrence. Time did not permit the determination of these endorsement distributions for the Summerfield sample.

In order to obtain that first-order factor structure which would be entirely based on built-in scale interdependency, the same procedures as outlined above were performed on the correlations derived from the 2500 randomly answered inventories. Ahmavaara's (1954) procedure for comparing factor structures was used to determine how similar the built-in factor structure was to the factor structure based on the responses of 400 Summerfield finalists.

² The results of the Cattell-White transformation and a third-order analysis can be obtained from the author by request.

Results

Ten alpha factors having positive generalizability were extracted from the "Summerfield" 60×60 intercorrelation matrix, accounting for 76.70 per cent of the original interscale variance. Table 1 presents the factor pattern matrix obtained from Maxplane. This solution was based on finding that reference vector structure which yielded a maximum .20 hyperplane ($\pm .10$) count. Out of 600 (60×10) reference vector coefficients, 64.2 per cent were .10 or less and 74.7 per cent were .20 or less in absolute value. Table 2 presents the intercorrelations among these 10 first-order alpha factors.

It is immediately evident from Table 1 that there are several instances in which a first-order factor is loaded on variables from different domains. However, on closer inspection of these loadings, there is a rather high degree of independence among the domains. In order to demonstrate this independence, the loadings above an absolute value of .40 have been italicized in Table 1. The figure is an arbitrary one, but the investigator believes this cut-off point provides a clear and meaningful presentation of the pattern of loadings, and yet is not so high as to delete loadings of practical significance in inferring the identity of a factor. Using this frame of reference, first-order factors IV, V, and VII are defined by MMPI variables, and the seven remaining factors are defined by SVIB variables.

It is also noticeable that the loadings on the two ability variables, CMT and SMT, do not achieve any appreciable magnitude ($\pm .40$) across all ten first-order factors, and therefore there is an absence of what might be termed an "ability factor." The communalities for these two variables were only .2497 and .1183, respectively, indicating that the major portion of their variance was specific (rather than shared) variance. Thus, a large portion of the variance of these two variables remained independent of the other two domains and unaccounted for by the common factor structure.

In Table 2 it can be seen that the analysis did result in correlated first-order factors with the exception of a few factor pairs which approach orthogonality. However, of the three "personality factors" (IV, V, and VII) only factor VII is substantially correlated with the seven "interest factors." Of the 10 first order factors, factor VII accounted for the least amount of variance. Factors IV and V are essentially orthogonal to these seven factors.

Four alpha factors were extracted from the correlations among the first-order factors. These four factors accounted for 56.85 per cent of the total variance expressed by the original 60×60 intercorrelation matrix. It should be noted that the final communality value for first-order factor VII was 1.12194, resulting in a Heywood case (Heywood, 1931). It was decided to allow this communality value to converge to a value greater than one on the basis of the advocacy of Kaiser and Caffrey (1965, pp. 8-9) who present the rationale that this procedure "implies only that the variable's unique part has negative variance—its unique factor scores are imaginary."

Table 3 presents the factor pattern matrix derived from the Max-plane rotation of the four second-order alpha factors. The percentage of variables (first-order factors) attained in the .20 hyperplane for this solution was 57.50.

Looking at the second-order factor pattern matrix as presented in Table 3, it is evident that in terms of factor composition the two areas of vocational interest and personality remain largely separate and distinct, with the exception of second-order factor III. Second-order factors I and II are entirely defined by first-order factors which have been previously defined as interest factors. Second-order factor IV is almost a complete reflection of first-order factor IV, a personality factor. Only second-order factor III has substantial loadings on first-order factors from both domains.

Space limitations made it necessary to condense the 58×58 matrix of intercorrelations based on 2500 randomly answered SVIB and MMPI inventories. Table 4 presents the frequency distributions of the built-in intercorrelations among the 45 scales of the SVIB. The interval size of .04 represents two standard errors of measurement (Guilford, 1965, p. 162).

Table 5 presents the intercorrelations among the *K*-corrected standard scores for the 13 basic MMPI scales that were based on 2500 randomly answered inventories.

Of the 585 (45×13) between-instrument correlations, only 9 (1.5 per cent) exceeded zero by more than two standard errors of measurement. The largest between-instrument correlation was .07.

Fourteen alpha factors were extracted from the 58×58 built-in intercorrelation matrix, accounting for 57.66 per cent of the original variance. Table 6 presents the factor pattern matrix obtained

TABLE 1

Factor Pattern Matrix for Ten First-Order Factors from Eber's Maxplane Oblique Solution Based on a Factor Analysis of the CMT, SMT, MMPI, and SVIB-M for 400 Summerfield Finalists

Group	Variable	I	II	III	IV	V	VI	VII	VIII	IX	X
I	Concept Mastery Test	-.32	.27				.19	.20	-.24	-.17	.22
	Stouffer Math Test	.30					.38				
	L					.50					
	F	-.15		-.27	.47	-.38			-.23		
	K			.15		.91					
	Hs				.49	.50		.73			
	D				.24	.21					
	Hy				.57	.54					
	Pd				.64	.20					
	Mf				.21	-.17		.20			-.37
	Pa	-.14			.54		.14	-.15			-.21
	Pt				.69			.22			
	Sc				.85				-.23		
	Ma				.61	-.44		-.60			
	Si		.18			-.32		.71			
	Artist		-.27	-.23							
	Psychologist	-.68		-.86							
	Architect	-.23	.95								
	Physician	-.30	.24	-.86			.44			.28	
	Osteopath		.73	-.32			.19			-.21	
	Dentist		.64							-.44	
II	Veterinarian			-.73		-.13				-.14	
	Mathematician	.23					-.36	.15		-.47	-.30
	Physicist	.32		-.79			.45				
	Chemist	.23		-.89			.34				.41
	Engineer	.20	.43	-.69			.30				.40
III	Production Manager	.33		-.83			.18				.64
		.47	.20					-.20			

TABLE 2

Intercorrelations of Ten First-Order Alpha Factors from Eber's Maxplane Oblique Solution Based on a Factor Analysis of the CMT, SMT, MMPI and SVIB-M for 400 Summerfield Finalists

	First-Order Alpha Factors								
	II	III	IV	V	VI	VII	VIII	IX	X
I	.37	-.38	-.08	-.05	-.50	.16	-.06	-.30	.07
II		-.53	.06	-.02	-.19	.32	-.37	-.54	-.61
III			.00	.02	.22	-.40	.21	.30	.23
IV				.02	-.09	.35	.05	.08	-.12
V					.09	-.30	.06	-.06	.03
VI						-.19	-.06	.37	.01
VII							-.26	.06	-.36
VIII								.26	.55
IX									.31

from the Maxplane solution. The percentage of variables attained in the .20 hyperplane was 76.35 per cent.

As would be expected, because of "chance" between-instrument correlations, the composition of each "built-in factor" is restricted to variables from either the SVIB or the MMPI. On inspection, many of the interest factors are predominately defined by a particular occupational grouping, e.g. factor III is defined by Group V, Social Service Occupations. With respect to the MMPI factors, one can justify the groupings of variables on each of these factors in terms of the variables sharing a K-correction, the sharing of a large number of items, or both (Dahlstrom and Welsh, 1960, p. 82).

TABLE 3

Factor Pattern Matrix for Four Second-Order Factors from Eber's Maxplane Oblique Solution Based on a Second-Order Analysis of the CMT, SMT, MMPI and SVIB-M for 400 Summerfield Finalists

First-Order Alpha Factor	I	II	III	IV
I	-.7752			
II	-.5752	.8998		
III	.4632	-.4127		
IV				.8088
V			.4274	
VI	.6405			
VII			-1.1101	
VIII		-.6441		
IX	.6493	-.6337	-.5263	
X		-.9045		

Note.—Loadings less than $\pm .40$ have been deleted for clarity of presentation.

TABLE 4

Distribution of Correlations among Scales of the SVIB Based on 2500 Randomly Answered Inventories

Interval	Frequency	Interval	Frequency
.77-.80	1	-.05--.08	60
.73-.76	1	-.09--.12	67
.69-.72	2	-.13--.16	49
.65-.68	1	-.17--.20	45
.61-.64	6	-.21--.24	42
.57-.60	8	-.25--.28	29
.53-.56	4	-.29--.32	27
.49-.52	13	-.33--.36	12
.45-.48	9	-.37--.40	7
.41-.44	26	-.41--.44	5
.37-.40	21	-.45--.48	2
.33-.36	28	-.49--.52	2
.29-.32	37	-.53--.56	0
.25-.28	45	-.57--.60	1
.21-.24	46	-.61--.64	2
.17-.20	59		—
.13-.16	51		
.09-.12	65		
.05-.08	57		
TOTAL	480	TOTAL	350

Note.—160 correlations did not differ significantly from zero at the .05 level (+.04 to -.04).

Table 7 presents Ahmavaara's transformation, comparing the built-in oblique first-order factors with the Summerfield oblique first-order factors.

From Table 7 the reader will notice that there are several instances where a built-in factor would be considered to be quite similar to that of a sample factor. These results suggest the inference that the

TABLE 5

TABLE 5
Correlations Between Scales of the MMPI Based on 2500 Randomly Answered Inventories

[illegible]

TABLE 6

Factor Pattern Matrix for Fourteen First-Order Factors from Eber's Marplane Solution Based on a Factor Analysis of the Obtained Built-in Matrix

Variable	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
L														
F														
K					.83									
Hs														
D														.61
Hy					-.71							.40		
Pd					.54									1.26
Mf														
Pa					.70					.62				
Pt					.47					.46				
Sc														
Ma								.59						
Si														
Artist														
Psychologist						.82								
Architect						.42								
Physician		.41		.51		.77							-.52	
Osteopath							.82							
Dentist							.94							
Veterinarian							.68							
Mathematician							.79							
Physicist						-.46							-.60	
Chemist		.40											-.59	
Engineer		.53											-.51	
Production Manager		.54												
Farmer	.42													
Carpenter	.83					-.50								
Forest Service Man	.46								.61					.42

TABLE 1—Continued

First-Order Factors

Variable	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
Aviator	.52					.42			.59					
Printer	.85					— .47							— .45	
Math Science Teacher	.51		.57											
Policeman	.71													
YMCA Physical Director	.45		.61											
Personnel Manager			.58											
Public Administrator			.58			.50			.61		— .40			
Social Worker			.77											
Social Science Teacher	.52		.80											
School Superintendent			.80											
Minister			.70											
Music Performer	.55					.72								
CPA Owner				.99		.66								
Senior CPA				.74										
Accountant				.83										
Office Worker				.68										
Purchasing Agent				.43										
Banker		— .43		.55							— .45			
Pharmacist							.65		— .44		— .64		— .56	
Mortician							.49				— .58			
Sales Manager											— .95			
Real Estate Salesman		— .47									— .64			
Life Insurance Salesman											— .57			
Advertising Man						.98					— .76			
Lawyer						.67								
Author-Journalist		— .46				1.02								
President MFG Concern	— .48										— .53			
IM			.48											.49
OL	— 1.06					— .43		.45						
MF								.55						

Note.—Loadings less than $\pm .40$ have been deleted for clarity of presentation.

TABLE 7

*Ahmavaara's Transformation Comparing Built-in-Factors with Sample-Factors
Based on Oblique Factor Pattern Matrices*

Built-in First-Order Factors	Sample-First-Order Factors									
	I	II	III	IV	V	VI	VII	VIII	IX	X
I	.41	.44	.00	-.01	.08	-.64	-.02	.18	.27	.09
II	.52	.63	-.45	.02	.09	.24	-.42	-.02	.43	.44
III	.14	.78	1.08	-.07	.04	.25	-.06	-.02	.03	-.06
IV	.82	.11	.70	-.03	-.03	.11	.07	.11	.41	.20
V	-.06	.15	.14	.60	.74	-.02	-.14	-.11	-.07	.23
VI	-.68	.58	-.38	.03	.05	.23	-.00	-.11	-.03	-.09
VII	-.07	.55	-.31	-.05	.00	-.03	.08	.68	-.41	.06
VIII	-.10	-.07	-.04	.95	-.30	.15	-.16	-.23	.06	-.13
IX	.21	.29	-.47	.09	-.10	.02	-.01	-.26	-.26	.92
X	-.04	.10	-.11	.92	-.29	-.06	-.06	-.16	.14	-.06
XI	.24	.10	-.60	-.09	.00	.24	.05	-.11	-.08	-.35
XII	.03	-.23	-.10	.16	-.24	.02	.84	-.04	.01	-.02
XIII	-.23	.05	.92	.04	.16	-.41	-.12	-.03	.01	-.09
XIV	-.04	.16	.02	.59	.77	.02	.00	.01	.00	.02

Note.—The coefficients in this comparison matrix represent the *loading* of a built-in factor on a sample factor, where the former has been transformed into the common factor space of the latter.

factor structure of "pooled" MMPI and SVIB variables may be largely built-in, in that several quite similar factors emerge even when these two inventories are responded to in a completely random manner.

Conclusions

The general conclusion of this study, disregarding the issue of interdependent scales, would be that the psychological realms of vocational interest (SVIB), personality (MMPI), and ability (CMT and SMT) have again been demonstrated to be, to a surprising extent, independent and unrelated. Each of the first-order factors was defined largely by a single domain, i.e. by either interest or personality variables. The two ability variables proved to be not only independent of the other domains but also of each other. At the second-order level there was only one factor which loaded on both interest and personality first-order factors. The writer is of the opinion that the results of this study do not provide enough evidence to support the previously described hierarchical trait model with any substantial degree of confidence.

When the possible implications of the effect of scale interdependency are considered, one is left with uncertainties and questions

rather than conclusions. It would be erroneous to conclude that, because of the demonstrated similarity between the sample structure and the built-in structure, the sample structure is largely a reflection of a methodological artifact rather than stable personality and interest dimensions. The opposite conclusion could be as easily reached. For example, in the case of the MMPI overlapping items were introduced because they discriminated between a normal group and two (or more) maladjusted groups. Therefore, overlapping items between two scales can be viewed as a possible measure of a dimension which is common to the two maladjusted groups in question, but which is not common to "normals." Might not these psychological dimensions be of a more fundamental and stable nature than those dimensions which represent the differentiation of only a single maladjusted group from normals? If one is to consider the built-in relationships among scales as completely meaningless for a given sample of subjects, then the assumption has to be made that the ability of overlapping items to discriminate between "normal" and "maladjusted" subjects is not valid for that particular sample of subjects. Is there any reason to suspect an overlapping item to be less valid across populations than a nonoverlapping item?

Although no definitive conclusions can be reached regarding the validity of built-in relationships for a given sample, the degree to which reported correlations among the scales of the MMPI and SVIB may be erroneous indices of relationship should at least be recognized. One might also speculate that the built-in relationships would have been even greater, and the similarities between the built-in structure and the sample structure more numerous, if the built-in correlations had been calculated on the basis of the distribution of endorsements given by the Summerfield sample, rather than assuming an equal probability of endorsement for each item alternative.

REFERENCES

- Ahmavaara, Y. Transformation analysis of factorial data. *Annales Academic Scintrarum Fennice*, Helsinki: Ser. B, 1954, 88(2), 1-150.
- Anderson, W. and Anker, J. Factor analysis of MMPI and SVIB scores for a psychiatric population. *Psychological Reports*, 1964, 15, 715-719.
- Bargmann, R. The statistical significance of simple structure in

- factor analysis. Frankfurt-Main: Hochschule fuer Internationale Pädagogische Forschung, 1953.
- Bartlett, M. S. Test of significance in factor analysis. *British Journal of Psychology*, Statistical Section, 1950, 3, 77-85.
- Becker, J. H. An exploratory factor analytic study of interests, intelligence, and personality. *Psychological Reports*, 1963, 13, 847-851.
- Bendig, A. W. and Meyer, W. J. The factorial structure of the scales of the Primary Mental Abilities, Guilford Zimmerman Temperament Survey, and Kuder Preference Record. *The Journal of General Psychology*, 1963, 68, 195-201.
- Cattell, R. B. The meaning and strategic use of factor analysis. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology*. Chicago: Rand McNally, 1966.
- Coan, R. W. Facts, factors, and artifacts: The quest for psychological meaning. *Psychological Review*, 1964, 71, 123-140.
- Dahlstrom, W. G. and Welsh, G. S. *An MMPI handbook*. Minneapolis: University of Minnesota Press, 1960.
- Eber, H. W. Toward oblique simple structure: Maxplane. *Multivariate Behavioral Research*, 1966, 1, 112-125.
- Guilford, J. P. When not to factor analyze. *Psychological Bulletin*, 1952, 49, 26-37.
- Guilford, J. P. *Fundamental statistics in psychology and education*. (4th ed.) New York: McGraw-Hill, 1965.
- Havlicek, L. L. *An investigation of the correlations between scales on the Strong Vocational Interest Blank*. (Doctoral dissertation, University of Kansas), Ann Arbor, Michigan: University Microfilms, 1965. No. 65-11, 928.
- Heywood, H. B. On finite sequences of real numbers. *Royal Society Mathematical and Physical Sciences Proceedings*, 1931, 134, 486-501.
- Kaiser, H. F. and Caffrey, J. Alpha factor analysis. *Psychometrika*, 1965, 30, 1-14.
- Royce, J. R. Factors as theoretical constructs. *American Psychologist*, 1963, 18, 522-528.
- Shure, G. H. and Rogers, M. S. Note of caution on the factor analysis of the MMPI. *Psychological Bulletin*, 1965, 63, 14-18.

TYPING SHIPS WITH TRANSPOSE FACTOR ANALYSIS

WILSON H. GUERTIN

University of Florida

CONVENTIONAL factor analysis examines the matrix of intercorrelations among tests. When the score matrix is oriented conventionally with people as rows and test variables as columns, the intercorrelations are computed between all possible pairs of columns.

Stephenson (1936) and Burt (1937) both are credited with recognizing the possibility of transposing the data matrix before intercorrelating columns and factor analyzing. When the transposed rows become columns, the intercorrelation of columns is an intercorrelation of people.

Factor analysis of the intercorrelations of test variables explicates clusters of these variables in test-space. When the correlation is between people the analysis is of people-space instead of test-space. Therefore, we can say that the transpose factor analysis explicates clusters of these persons in people-space.

With superficial classification the correspondence between person type and trait characterization may be high. For example, everybody who is very high on dishonesty we could call a "crook." In such a system the classification of crook or noncrook can be made from a knowledge of a person's one attribute, like honesty. If dishonesty and intelligence enter the classification scheme as dichotomous determinants there are four classes but only two trait measures. With multivariate trait bases for classification the number of type-factors and trait-factors will not necessarily be equal nor will there be an obvious correspondence between two kinds of factors.

Burt (1940) exaggerated the correspondence between the conventional and transpose factor analyses. He placed such conditions

on the correlation matrices that his correspondence is obviously mathematical and of theoretical rather than practical importance. The conditions are that: raw scores be normalized and ipsatized, cross-products be used instead of correlations, unities be placed in the diagonal, and finally, no rotation of the factor matrix be made. Such restrictions as these take the analysis far from the realm of common-factor analysis and the practical concerns of the data analyst.

A comparison of the two sets of obtained factors is illustrated for geometric objects in a study by Lorr, Jenkins, and Medland (1955). Even with only four attribute (trait) factors they found only moderate correspondence between sets of factors. Cattell probably has been over-reacting to Stephenson's extravagant claims for transpose analysis (called Q-technique by Cattell as contrasted with the conventional or R-technique) and Q-sorting. However, Cattell states now "... no simple equivalence of meaningful factors from ordinary R- and Q-technique results." (1966, p. 228).

Cattell's reaction to Stephenson's claims may well be part of the basis for his continued ambivalence toward transpose factor analysis as a means of deriving taxonomic principles. Before Stephenson's book appeared (1953) Cattell wrote, "Q-technique is most useful if one wishes immediately to see how many types there are in a population and to divide it up into types." (1952, p. 101). Cattell continues to employ parametric statistics in conventional analyses but reverts to nonparametric cluster analysis when seeking taxonomic principles in the interperson similarities matrix derived from distances in the transposed data matrix. He now goes so far as to deny that types are factors! (Cattell and Coulter, 1966, p. 242). Much of Cattell's disappointment may have come from using correlation earlier as the index of similarity between people.

Cronbach and Gleser (1953) pointed out that the product-moment r fails to express differences in the average levels (means) as well as differences in average scatter (variance) of two profiles. They recommend use of the distance statistic, d . This d is the square root of the sum of the squared differences between two profiles across all variables.

Factor analyses have been made of covariance matrices (Horst,

1965) and cross-product matrices (Nunnally, 1962) but these produce very different results than those obtained by factor analyzing interprofile d 's. Use of d instead of covariance or cross-products permits the extraction of common factors. The only d factor analyses reported to date are in connection with Guertin's Successive Profile Analysis procedure (1966). It is time a clearcut illustration of typing with factor analysis of interprofile d 's appeared in the literature.

Transpose Analysis of Ship Intercorrelations

Cattell and Coulter provided the ship data they analyzed with their "taxonome" procedure (1966, p. 265). There were 12 measures on each of 29 vessels from data in Jane's Fighting Ships (1964-65). The object of the task is to find out how many types of ships there are and which ships belong to each class.

All measurements were put in standard score form and are as follows: displacement, length, beam, number of light, of medium, of heavy, and of very heavy guns, number of personnel, maximum speed, submersibility, (obviously dichotomous), continuity of deck construction, and number of planes carried.

If we intercorrelate all the 29 ships across the 12 measures, then those ships which are alike because they are of the same type should be highly intercorrelated. Inspection of the matrix of correlations should disclose four clusters or submatrices of high intercorrelations because, as we find out later, there are four types of ships. Each cluster constitutes a group of ships that should be alike (in a correlational sense) but bear little "similarity" to members of other clusters. The factors produced can be viewed as possible type categories for the classification job at hand.

Principal axes from the reduced intercorrelation matrix of the 29 ships were rotated to give the Varimax rotated factors which appear in Table 1. Loadings greater than .49 are italicized for emphasis. The five carriers are clearly identified as belonging together in a class by themselves by having positive loadings on factor 1. The submarines are negatively related to the class and all have negative loadings of at least .40. Thus, the first factor identifies another type of ship at the negative end.

The second factor gives perfect identification of the destroyer type and its members. Again submarines are loaded negatively.

TABLE 1

Varimax Rotated Matrix for Correlational Analysis of Ships Problem

Ship		Factor			
		I	II	III	IV
Carrier	1	.91	.10	-.18	-.29
	2	.85	.36	-.30	-.09
	3	.94	.09	-.11	-.23
	4	.95	.11	-.10	-.27
	5	.83	.33	-.28	-.19
Destroyer	1	.22	.94	.09	-.03
	2	.24	.96	.02	-.06
	3	.12	.96	.17	-.09
	4	.03	.96	.09	.00
Submarine	1	-.43	-.52	-.55	.42
	2	-.48	-.55	-.52	.41
	3	-.59	-.71	-.29	-.12
	4	-.55	-.48	-.22	-.60
	5	-.54	-.58	-.52	-.22
	6	-.64	-.56	-.28	-.28
	7	-.62	-.56	-.23	-.33
	8	-.40	-.50	-.26	-.22
	9	-.41	-.50	-.56	-.42
	10	-.55	-.63	-.40	.00
Frigate	1	-.14	.26	.82	.44
	2	-.14	.11	.93	.28
	3	-.11	.19	.86	.39
	4	-.15	.07	.92	.26
	5	-.18	-.10	.50	.80
	6	-.16	-.20	.46	.80
	7	-.16	-.05	.17	.96
	8	-.17	.15	.18	.94
	9	-.07	.24	.94	.13
	10	-.15	.06	.36	.91

Using the cutting line, .49, only submarine number four lies outside the class with a loading of $-.48$.

The third factor centers around the frigates but some class members load weakly. Nor is this a case of factor fission where variance from the frigate cluster is pulled out into another dimension because too many factors were rotated. We can be sure of this because the R matrix (not reported here) shows frigate 4 negatively correlated with frigates 7 and 8. The fourth factor behaves much like the third in getting only half of the frigates to load heavily.

The overall results stack up to indicate the two clear types of carrier and destroyer with unequivocal identification of the members of each. Submarines are less homogeneous but are recognizable as a type or two subtypes. The frigates break into two dis-

tinct types. The results of the analysis are by no means satisfactory so we are led to consider other possibilities for analysis, namely, factor analyzing *d*.

Transpose Analysis of Ship Distances

The existing concepts for factor-space based upon a correlation matrix are not appropriately named for dealing with factor-space derived from a distance-based index of similarity. A distinction must be made between test-space (Thurstone's term which in more general form would be attribute-space) and person-space. On the other hand a distinction must be made between both test- or person-space based upon relation indices and that from distance indices. When space is used as the analogue of relationship the term relation-space is appropriate. When space is the analogue

TABLE 2

Varimax Rotated Matrix for Distance Analysis of Ships Problem

Ship		Factor			
		I	II	III	IV
Carrier	1	.78	.27	.26	.24
	2	.82	.23	-.02	.00
	3	.89	.11	.01	.03
	4	.89	.20	.14	.14
	5	.72	.38	.27	.25
Destroyer	1	.39	.80	.12	.24
	2	.39	.81	.19	.29
	3	.32	.78	.23	.37
	4	.27	.82	.15	.30
Submarine	1	.17	.15	.91	.22
	2	.15	.14	.91	.24
	3	.10	.10	.84	.36
	4	.15	.18	.83	.25
	5	.07	.08	.89	.22
	6	.01	.09	.86	.23
	7	.05	.11	.85	.30
	8	.10	.12	.70	.45
	9	.17	.15	.89	.20
	10	.00	.05	.78	.36
Frigate	1	.15	.25	.34	.85
	2	.14	.25	.27	.85
	3	.15	.24	.35	.85
	4	.14	.24	.30	.84
	5	.05	.15	.20	.78
	6	.08	.13	.31	.82
	7	.09	.16	.35	.78
	8	.13	.21	.41	.78
	9	.17	.28	.34	.81
	10	.13	.19	.42	.79

for representing distances between people or tests the term distance-space is appropriate.

The statistic d is a distance-based measure of dissimilarity so it is necessary to transform it to a similarity index. The d 's can be changed to similarity indices by simply reflecting them, i.e. subtract them from a constant. The arbitrary constant from which each is subtracted is not critical: we use the largest d in the off-diagonal elements.

Subtracting a distance measure from the largest value present is to employ a very arbitrary reference value. The largest d value will depend upon how far apart the two most dissimilar profiles are. While we cannot eliminate the noncomparability of the magnitude of d between studies, we can eliminate the effect from the interprofile similarity matrix. This will be done for the examples used here by dividing each element of the reflected matrix by the largest element in it. Thus, the magnitudes of the index range from 0 to 1, and the dependence of the index on the arbitrary magnitude of score scales is eliminated. The result is a new index and we refer to it as the *distance similarity index* or DSI.

The concept of common person-distance factor-space appears sound. Person-distance communality can be estimated by taking the largest column index value and factoring iteratively to get a final value, thus supporting the analogy between the two types of factor-space. The correlation index of interprofile similarity gave type-factors with uncertain correspondence to actual classes of vessels. What would a factor analysis of interprofile d 's show? Table 2 gives the Varimax rotated matrix for the analysis. The communality estimate employed was the largest value in the column. It has been demonstrated (Guertin, 1969) that such estimates are satisfactory.

This time the rotated matrix unequivocally indicates four classes of vessels. The italicized "loadings" are those above .49. None of these appreciable loadings is misplaced and none needed to indicate true class membership is missing.

Illustration of Computations

To remove any ambiguity about computational procedure some of the calculations will be presented here. It will suffice to work with the upper left submatrix (first three profiles).

TABLE 3
First Three Profile Submatrices

Variables	Ships			Sum d^2		
	1	2	3	1	2	3
1	1.50330	1.84090	1.81500	1	14.2286	10.3423
2	1.12770	2.82440	2.88780	2		4.2875
3	1.45080	2.31550	2.31550	3		
4	.84196	1.88100	1.88100			
5	1.77480	2.07140	2.56170		Square Root Sum d^2	
6	-.09409	2.06010	.04952	1	2	3
7	-.50175	.72136	.72136	1	3.7721	3.2159
8	-.63840	-.63840	-.63840	2		2.0706
9	1.22420	2.89220	2.89220	3		
10	.53551	1.16940	1.16940			
11	-.72548	-.72548	-.72548		BIG- d	
12	-.63830	-.63830	-.63830	1	2	3
			1		5.6047	6.1609
			2			7.3062
			3			
					d/BIG' or DSI	
				1	2	3
			1	.8781	.6118	.6725
			2		.7975	.7975
			3			.8353

Table 3 gives the successive submatrices from standard scores to the final similarity index values. The sum d^2 values are derived from the sum of the squared differences between pairs of scores for ships on all 12 variables. The next step is to take the square root. The index for 1-2 is 3.7721 and the largest in the submatrix but the largest value in the complete matrix (not shown here) is that between 2 and 19. It is labelled "BIG" and is equal to 9.3768. Computations have been carried to 10 decimal place accuracy but rounded to four places for presentation in the tables.

Next, all d 's are subtracted from BIG. While 2-3 is the largest in the resulting submatrix with a value of 7.3062, that in the complete matrix (not shown here) is 9.1910 for ships 20-22, $\text{BIG}' = 9.1610$ is used to divide each element so these final similarity index values range from 0.00 to 1.00. It is this last matrix with communality estimates inserted in the diagonal which is factor analyzed.

Summary

The value of distance indices for profile matching is discussed in the literature occasionally and demonstrated even less frequently. Similarly, transpose factor analysis has been the subject

of hot debate but little action. The value of these procedures in identifying classes for the assignment of individuals with recurring similar profiles by analyzing their profiles needs to be illustrated.

Cattell supplied 12 measures on each of 29 fighting ships. Each ship was known to be a carrier, destroyer, submarine, or frigate. Intercorrelation of the transposed score matrix with multiple correlation estimates of communality in the diagonal was factored and rotated to a Varimax solution. It failed to clearly identify the number of classes and which ships belonged in each.

Distance measures between profiles were computed, reflected to represent similarities, and then divided by the maximum value in the interprofile matrix of similarities. The values in this final matrix which was factor analyzed must necessarily range from 0.00 to 1.00. It is proposed that this index based upon d be called the *distance similarity index* and abbreviated as DSI.

The transpose factor analysis of the DSI interprofile matrix was much more successful than the transpose factor analysis of the intercorrelations of profiles. In fact, the DSI analysis produced the four type-factors corresponding to the four ship-classes. Furthermore, each ship was unequivocally classified correctly by taking the highest of the four type-loadings as the basis.

REFERENCES

- Burt, C. L. Correlations between persons. *British Journal of Psychology*, 1937, 28, 59-96.
- Burt, C. L. *Factors of the mind*. London: University of London Press, 1940.
- Cattell, R. B. *Factor analysis. An introduction and manual for the psychologist and social scientist*. N. Y.: Harper, 1952.
- Cattell, R. B. (Ed.) *Handbook of multivariate experimental psychology*. Chicago: Rand McNally, 1966.
- Cattell, R. B. and Coulter, M. A. Principles of behavioral taxonomy and the mathematical basis of the taxonome computer program. *British Journal of Mathematics and Statistical Psychology*, 1966, 19, 237-269.
- Cronbach, L. J. and Gleser, G. C. Assessing similarity between profiles. *Psychological Bulletin*, 1953, 50, 456-473.
- Guertin, W. H. The search for recurring patterns among individual profiles. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1966, 26, 151-165.
- Guertin, W. H. Factor analysis of interprofile distances. Paper read at AERA, Los Angeles, 1969.

- Horst, P. *Factor analysis of data matrices*. N. Y.: Holt, Rinehart & Winston, 1965.
- Lorr, M., Jenkins, R. L., and Medland, F. F. Direct versus obverse factor analysis: A comparison of results. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1955, 15, 441-449.
- Nunnally, J. The analysis of profile data. *Psychological Bulletin*, 1962, 59, 311-319.
- Saunders, D. R. and Schueman, H. Syndrome analysis: An efficient procedure for isolating meaningful subgroups in a non-random sample of a population. Paper read at Psychonomic Society, St. Louis, 1962.
- Stephenson, W. The inverted factor technique. *British Journal of Psychology*, 1936, 26, 344-361.
- Stephenson, W. *The study of behavior*. Chicago: University of Chicago Press, 1953.

CONSIDERATIONS WHEN MAKING INFERENCES WITHIN THE ANALYSIS OF COVARIANCE MODEL¹

CHARLES E. WERTS AND ROBERT L. LINN

IN his discussion of multiple regression, Cohen (1968) notes that the analysis of covariance (ANCOVA) is equivalent to a regression analysis in which treatment groups are coded as dummy variables. The numerator of the traditional ANCOVA F test (McNemar, 1962) is equivalent to the squared total multiple correlation of dummy variables and covariates with the dependent variable minus the squared multiple correlation of the covariates with the dependent variable and the denominator of the F test is equivalent to the error variance, i.e., one minus the squared total multiple correlation. The numerator is therefore what Darlington (1968) calls "usefulness," i.e., the proportion of variance that the dummy variables add to the prediction of the dependent variables in a stepwise regression after the covariates have entered. However, Linn and Werts (1969) point out that "usefulness" requires a number of assumptions when used in making causal inferences. It follows that the considerations listed by Linn and Werts (1969) can be specialized to the analysis of covariance method as shall be demonstrated below.

The ANCOVA Model

For purposes of discussion consider the case of a single covariate (X), three treatment groups ($j = 1, 2, 3$), and a dependent variable, Y . The mathematical model for ANCOVA is then:

$$Y_{ij} = A_i + B_w X_{ij} + e_{ij} \quad (1)$$

¹ The research reported herein was performed pursuant to Grant No. OEG-1-6-061830-0650 Project No. 6-1830 with the Office of Education, U. S. Department of Health, Education and Welfare.

where

A_j = the Y intercept of the Y on X regression line for group j .

Also

$A_j = \bar{Y}_j - B_w \bar{X}_j$ (\bar{X}_j and \bar{Y}_j are the respective means of X and Y for group j),

B_w = pooled within group regression slope, and

e_{ij} = error term assumed to be independent of A_j and X_{ij} and with zero mean.

In addition to the usual linear regression assumptions, this model requires homogeneity of within-group regression. The traditional ANCOVA procedure also requires that the treatment effect (i.e., A_j) and the covariate (X_{ij}) be independent (Evans and Anastasio, 1968). Furthermore, it must follow from substantive theoretical considerations that the treatment means *should* be adjusted (Smith, 1957) using the within-group regression slope (i.e., $A_j = \bar{Y}_j - B_w \bar{X}_j$). Providing reasonable justification for the use of ANCOVA in non-experimental situations may be quite difficult as Lord (1969) clearly demonstrates. Because of this it is frequently necessary to abandon strong causal interpretation and to develop insights that establish models in which causal inferences can be tested later.

Following Cohen's (1968) procedure, equation (1) can be translated for computational purposes to the dummy variable form:

$$Y_{ij} = B_1 Z_1 + B_2 Z_2 + B_3 Z_3 + B_w X_{ij} + e_{ij} \quad (2)$$

where the dummy variable coding is:

$Z_1 = 1$ for everybody

$Z_2 = 1$ for persons in group #2, 0 for others, everybody,

$Z_3 = 1$ for persons in group #3, 0 for others.

When equation (2) is solved in the usual regression program the A_j intercepts in equation (1) can be computed directly from the regression weights in equation (2) as shown by Johnston (1960), i.e., $A_1 = B_1$, $A_2 = B_1 + B_2$, and $A_3 = B_1 + B_3$. Since Cohen's procedure involves the computation of correlations, it is useful to convert equation (2) to its standardized form:

$$y_{ij} = b_2^* z_2 + b_3^* z_3 + b_w^* x_{ij} + e_{ij}, \quad (3)$$

where lower case letters indicate standardized variables and b_i^* are

the standardized regression weights. b_1^* is zero since Z_1 in equation (2) is a constant.

The "usefulness" of the dummy variables in terms of the total multiple correlation $R_{y \cdot z_2, z_3, x}$ is then:²

$$R_{y(z_2, z_3, x)}^2 = \text{"usefulness"} = R_{y \cdot z_2, z_3, x}^2 - R_{yx}^2$$

where $R_{y(z_2, z_3, x)}$ is the multiple part correlation of the dummy variables with the dependent variable, covariate partialled out of the dummy variables. The subscript z_2, x indicates the residual of z_2 with x controlled and z_3, x is the residual for z_3 with x controlled. As should be suspected from a comparison of equations (1), (2), and (3) the multiple part correlation $R_{y(z_2, z_3, x)}$ is equal to the part correlation $R_{(A \cdot X)Y}$, i.e., the correlation of the dependent variable with the intercepts when the covariate is partialled out of the intercepts. In other words, we could have taken the intercepts calculated in the dummy variable analysis, inserted these in equation (1), and computed the "usefulness" or "adjusted treatment variance" from the correlations among X_{ij} , Y_{ij} and A_{ij} :

$$R_{(A \cdot X)Y}^2 = R_{Y \cdot AX}^2 - R_{YX}^2 = \frac{(R_{AY} - R_{AX}R_{YX})^2}{(1 - R_{AX}^2)} = \text{"usefulness."}$$

This leads to the useful interpretation that ANCOVA and its dummy variable regression form is equivalent to assigning each person in a group the value of the "effect" (i.e., A_{ij}) for that group, this effect being scaled in units of the dependent variable. One may then think in terms of a single "treatment" variable which is one of the independent variables in the regression equation with the covariates as the remaining independent variables, though considerations about degrees of freedom are a little more complicated this way.

Application of Alternate Procedures

The essence of the above argument is that a categorical variable (treatment groups) may for analytical purposes be replaced by a

² The notation for the multiple partial beta is our own, since we know of no references to such a coefficient other than Werts (1968). Paralleling DuBois' (1957) notation, the first letter of the subscript refers to the dependent variable, the letters in the parentheses refer to the relevant independent variables, i.e., the dummy variables z_2 and z_3 , and the letter after the dot is the controlled variable.

single variable representing the treatment "effect" on the dependent variable. According to ANCOVA the proportion of variance in the dependent variable that this treatment variable "accounts for" is the squared part correlation $R_{(A.X)Y}^2$, however, several other procedures exist for calculating the proportion of "unique" variance in the dependent variable that a variable "accounts for." In addition to $R_{(A.X)Y}^2$, the following procedures could be and have been employed in regression analysis:

- a. The multiple partial correlation $R_{(y.z)(s_2.s_3.s_4)}$ of the dummy variables with the dependent variable

$$R_{(y.z)(s_2.s_3.s_4)}^2 = R_{YA.X}^2 = \frac{(R_{AY} - R_{AX}R_{YX})^2}{(1 - R_{AX}^2)(1 - R_{YX}^2)}$$

- b. The multiple part correlation $r_{(y.z)(s_2.s_3)}$ of the dummy variable with the dependent variable, covariate(s) partialled out of the dependent variable:

$$R_{(y.z)(s_2.s_3)}^2 = R_{A(Y.X)}^2 = \frac{(R_{AY} - R_{AX}R_{YX})^2}{(1 - R_{YX}^2)}$$

- c. The multiple standardized partial regression coefficient $b_{y(s_2.s_3).x}^*$ of the dummy variable with the dependent variable when the covariates are also independent variables:²

$$\begin{aligned} (b_{y(s_2.s_3).x}^*)^2 &= (b_{YA.X}^*)^2 = (b_2^*)^2 + (b_3^*)^2 + 2b_2^*b_3^*r_{23} \\ &= \frac{(R_{AY} - R_{AX}R_{YX})^2}{(1 - R_{AX}^2)^2} \end{aligned}$$

where r_{23} is the correlation of z_2 and z_3 .

Thus four techniques for obtaining the proportion of variance that the categorical treatment variable "accounts for" are possible in a regression analysis:

- Method 1. The squared partial correlation $R_{YA.X}^2$
- Method 2. The squared part correlation $R_{(A.X)Y}^2$
- Method 3. The squared part correlation $R_{A(Y.X)}^2$
- Method 4. The squared beta weight $(b_{YA.X}^*)^2$.

Insofar as the researcher wishes only to test the null hypothesis of no improved predictability effect, it does not matter which approach is used since the same F ratio and degrees of freedom for testing statistical significance is applicable to all four methods. The

F test of the unstandardized regression weight $B_{YA.X}$ [equals unity as seen in equation (1)] also yields the identical level of significance. If the researcher wishes to interpret the magnitude of the "proportion of variance accounted for" as indicating a large or small treatment effect, then the theoretical considerations detailed by Linn and Werts (1969) become relevant, for example, Method 3 ($R_{A(Y.X)}$) requires the assumption that the "treatment effect" and the covariate be independent. Method 4 ($b_{YA.X}^*$) is the only one of the four methods which allows for X and A to be nonindependent determinants of Y . Even for Method 4, however, a strong model must be adopted to enable the regression weight to be interpreted as a "treatment effect."

Implications for Understanding and Extending ANCOVA

The foregoing discussion indicates that given the assumption of within-group homogeneity of regression the ANCOVA model is a special case of the general linear regression model. To understand the traditional ANCOVA terminology in terms of this regression approach, consider Figure 1 which is a diagrammatic representation of the model $Y_{ij} = A_j + b_w X_{ij} + e_{ij}$.

In this model the covariate and the treatment variable are determinants of the dependent variable Y as indicated by the respective arrows. The double headed arrow between A_j and X_{ij} indicates a correlation and the isolated arrow from e_{ij} to Y represents the

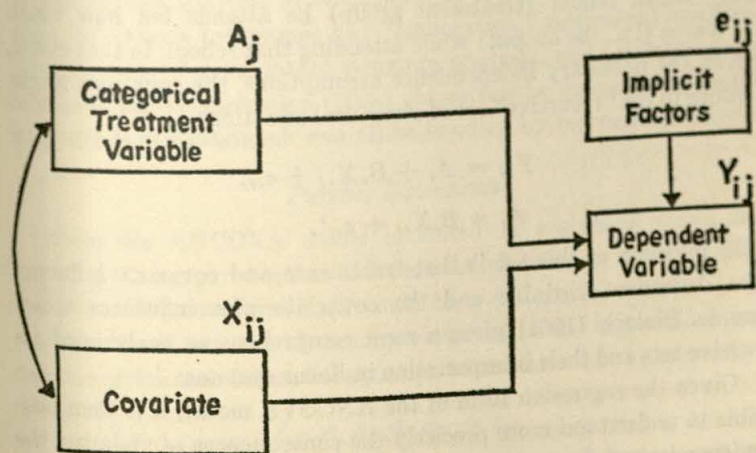


Figure 1. Representation of ANCOVA Model.

assumption that the arrows are assumed independent of A_i and X_{ij} or in causal language that all other (implicit) unmeasured factors influencing Y are assumed to be independent of both the covariate and the treatment variable.

The ANCOVA assumption that treatments and covariate be independent amounts to the assertion that any correlation between A_i and X_{ij} will arise solely from sampling errors which is equivalent to asserting that the between groups or external slope B_{FX} will differ from the within groups or internal slope B_w only because of sampling errors. Because of this assumption ANCOVA is seldom applicable to naturalistic studies where treatment group and covariate are usually associated because of systematic, nonrandom influences. However, techniques have been worked out within the framework of regression analysis for dealing with independent variables which are systematically related; discussions of these techniques can be found in a number of sources such as Turner and Stevens (1959), Tukey (1954), Blalock (1961), and Duncan (1966).

It is preferable to adopt a regression approach rather than pretend to adopt ANCOVA when its assumptions are violated. It also facilitates thinking in terms of a model of causality that explicitly allows for an association between the covariate on the treatment. Suppose that the covariate influenced which treatment group a subject was assigned to, e.g., it seems likely that a student's family background (school input or covariate) influences not only which school (treatment group) he attends but how much he learns (i.e., the output) while attending that school. In that event, given the necessary independence assumptions, the equations might reduce to the "recursive" set of structural equations:

$$Y_{ij} = A_i + B_w X_{ij} + e_{ij},$$

$$A_i = B_x X_{ii} + e_{ii}'.$$

The meaning of this set is that treatments and covariate influence the dependent variable and the covariate also influences treatments. Blalock (1961) gives a most comprehensive analysis of recursive sets and their interpretation in linear systems.

Given the regression form of the ANCOVA model, it is then possible to understand more precisely the consequences of violating the independence assumption (Evans and Anastasio, 1968). In this re-

gard, Winer (1962) states: "When the covariate is actually affected by the treatment, the adjustment process removes more than an error component from the criterion, it also removes part of the treatment effect." In more detail, it follows from the principles of regression analysis that when there are two independent variables, i.e., A_j and X_{ij} , then removing from the predictable variance all variance associated with one of them, i.e., X_{ij} , will necessarily remove that part of the variance ascribable to the other variable, i.e., A_j , which is correlated with the first variable. Thus if A_j and X_{ij} are correlated r_{AX} because of nonindependence, the variance predictable from X_{ij} will include $r_{AX}^2 \sigma_A^2$ of the variance in A_j . The adjusted treatment variance in ANCOVA will therefore be too small (i.e., biased) by the factor $(1 - r_{AX}^2)$ when treatment effect and covariate are nonindependent. In ANCOVA language, the adjusted treatment sums of squares (i.e., T_{YYR}) will equal the sums of squares (i.e., T_{YYA}) of the adjusted treatment means times the factor $(1 - r_{AX}^2)$. The usual formula relating the between-groups regression slope to the within-groups slope [e.g., equation (1) or (2) on p. 585 of Winer, 1962] can be converted to show the same relationship. In the presence of nonindependence the usual ANCOVA procedure is therefore biased, yielding an underestimate of the treatment variance. This problem is a well-known regression phenomenon, discussions of which may be found in most econometrics texts (e.g., Malinvaud, 1966). An important advantage of casting the analysis of covariance in a regression framework is that the various techniques like "elaboration," contextual analysis, ecological correlation, latent structure analysis, and Guttman scale analysis, which Schuessler (1968) shows can be expressed in terms of analysis of covariance, can all be handled by regression analysis.

Further Extensions

When the ANCOVA model is stated in regression terms the covariate is simply another independent variable in the regression equation. The covariate itself could be a categorical variable (e.g., B_K) expressed as a set of dummy variables in the regression equation; in which case the model equation becomes:

$$Y_{ij} = A_i + B_K + e_{ij}.$$

Johnston (1960) gives an excellent account of how to set up

the dummy variable analysis for this case and others. Note that this model equation is that of a two-way analysis of variance which, however, requires the assumption of independence of the two categorical variables A_j and B_K . The dummy variable analysis calculates least squares estimates of A_j and B_K even if A_j and B_K are nonindependent. Given A_j and B_K , the problem resolves itself into the familiar regression problem of two nonindependent continuous variables which can be analyzed with any appropriate regression technique mentioned earlier. That it is possible to analyze data cross-categorized in two or more ways does not, of course, mean that it is meaningful to do so.

Application to the Analysis of Compositional Effects

It is sometimes reasonable to hypothesize that gathering together a group of people with particular characteristics may result in a "compositional" effect on some outcome variable, an effect which is not predictable from the individual characteristics alone. If the group mean \bar{X}_i were the relevant indicator of composition, X_{ii} the individual characteristic, and Y_{ii} the outcome then in the regression model $Y_{ii} = B_0 + B_1 X_{ii} + B_2 \bar{X}_i + e_{ii}$ the partial regression weight for $\bar{X}_i (B_2)$ represents the net compositional influence with individual characteristics held constant and the weight for $X_{ii} (B_1)$ represents the net individual influence with composition controlled. A least squares solution yields $B_1 = B_W$ (i.e., the pooled within groups slope) and $B_2 = B_{FX} - B_W$ (i.e., the between minus the within groups slopes).

If treatments are independent of the covariate in ANCOVA the external slope (B_{FX}) equals the internal slope (B_W) which would mean that $B_2 = 0$ and that there would be no "compositional" effect. It can be shown that B_2 is the regression weight for A_i on \bar{X}_i as follows:

1. $A_i = \bar{Y}_i - B_W \bar{X}_i$.
2. Which has the normal equation

$$\sigma_{AX} = \sigma_{XF} - B_W \sigma_X^2.$$

3. Dividing by σ_X^2 yields

$$\frac{\sigma_{AX}}{\sigma_X^2} = \frac{\sigma_{XF}}{\sigma_X^2} - B_W,$$

or

$$B_{AX} = B_{FX} - B_W.$$

4. Therefore $B_1 = B_{1X} = B_{TX} - B_W$.

The relationship of the ANCOVA model to the "compositional" model is clarified further by noting that in the equation

$$Y_{ij} = B_0 + B_1X_{ij} + B_2\hat{X}_i + B_3A_i + \epsilon_{ij}$$

that $B_1 = B_W$, $B_2 =$ zero, and $B_3 =$ unity which means that this equation reduces to the ANCOVA model. The implication is that the "compositional" effect is in fact part of the "treatment" effect in ANCOVA.

To summarize: (a) the analysis of "compositional" effects corresponds to the ANCOVA model in which treatments are not independent of the covariate, which means that the assumptions of the ANCOVA model, especially within group homogeneity of regression apply also to compositional analysis; (b) the regression slope (B_{1X}) of the intercepts (i.e., A_i in the ANCOVA model) on the composition indicator (i.e., \hat{X}_i) represents the net influence of composition; and (c) the "compositional" effect is part of the "treatment" effect in the ANCOVA model.

Overview

It is well known that analysis of variance or analysis of covariance are simply specialized cases of the general linear equation, cases which are appropriate almost exclusively for experimental studies in which such assumptions as independence of covariate and treatments are plausible. Naturalistic and quasi-experimental studies, however, require analytical models which allow for the covariate to influence or be influenced by treatments or for nonindependent sets of categorical and/or continuous variables. Given modern computing facilities all analyses can be efficiently set up in a general linear format even if the requirements for ANOVA or ANCOVA are given. However, in the general linear model one is then faced with choosing among a variety of analytical procedures such as multiple partial correlations, multiple part correlations, and multiple partial regression weights (standardized and/or unstandardized). Which procedure to use and how to interpret the results provides the real test of a good researcher since good choices depend on an understanding of the phenomena being studied and of the relationship of the phenomena

to the mathematical model underlying the statistics. When the assumptions of an ANCOVA model are violated, causal inference must rest upon the assumption that you are dealing with a closed system or that all other variables that might influence the dependent variable are unrelated to the treatment or the covariate.

REFERENCES

- Blalock, H. M. *Causal inferences in non-experimental research*. Chapel Hill: The University of North Carolina Press, 1961.
- Cohen, J. Multiple regression as a general data-analytic system. *Psychological Bulletin*, 1968, 70, 426-443.
- Darlington, R. B. Multiple regression in research and practice. *Psychological Bulletin*, 1968, 69, 161-182.
- DuBois, P. H. *Multivariate correlational analysis*. New York: Harper & Brothers, 1957. Ch. 8.
- Duncan, O. D. Path analysis: Sociological examples. *The American Journal of Sociology*, 1966, 72, 1-16.
- Evans, S. H. and Anastasio, E. J. Misuse of analysis of covariance when treatment effect and covariate are confounded. *Psychological Bulletin*, 1968, 69, 225-234.
- Johnston, J. *Econometric methods*. New York: McGraw-Hill, 1960.
- Linn, R. L. and Werts, C. E. Assumptions in making causal inferences from part correlation, partial correlations and partial regression coefficients. *Psychological Bulletin*, 1969, 72, 307-310.
- Lord, F. M. Statistical adjustments when comparing preexisting groups. *Psychological Bulletin*, 1969, 72, 336-337.
- Malinvaud, E. *Statistical methods of econometrics*. Chicago: Rand McNally, 1966.
- McNemar, Q. *Psychological statistics*. New York: Wiley, 1962.
- Schuessler, K. Covariance analysis in sociological research. In E. F. Borgatta (Ed.), *Sociological Methodology*. York, Pennsylvania: Jossey-Bass, 1968.
- Smith, H. Interpretation of adjusted treatment means and regressions in analysis of covariance. *Biometrics*, 1957, 13, 282-308.
- Tukey, J. W. Causation, regression, and path analysis. In O. Kempthorne, T. A. Bancroft, J. W. Gowen, and J. L. Lush (Eds.), *Statistics and mathematics in biology*. Ames, Iowa: Iowa State College Press, 1954.
- Turner, M. E. and Stevens, C. E. The regression analysis of causal paths. *Biometrics*, 1959, 15, 236-258.
- Werts, C. E. The partitioning of variance in school effects studies. *American Educational Research Journal*, 1968, 5, 311-318.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

HOW TO WRITE TRUE-FALSE TEST ITEMS

ROBERT L. EBEL
Michigan State University

Why Use True-False Items?

The basic reason for using true-false test items is that they provide a simple and direct means of measuring the essential outcome of formal education, which is command of useful verbal knowledge. For all knowledge can be expressed in a series of propositions, and a proposition is simply a sentence that can be said to be true or false. Propositions are the substance of knowledge. Judging their truth or falsity is the essential task of scholarship in any field.

Some test constructors obtain scores of respectable reliability from true-false classroom tests. Examples are given in Table 1. These reliabilities indicate that the items in these tests were not seriously ambiguous, and that guessing could not have been extensive. If an ambiguous true-false item is written, it is the fault of the writer, not of the form. Also, when guessing affects test scores seriously, it is because the test is too short, the items too difficult or too ambiguous or the examinees too poorly motivated.

Compared with other item forms, true-false test items are relatively easy to write. They are simple declarative sentences of the

TABLE 1
Reliability of True-False Tests

Date	Students	Items	Reliability
3-6-68	114	110	.85
4-29-68	100	111	.77
5-27-68	100	107	.83
7-8-68	142	99	.86
7-23-68	141	90	.78

kind that make up most oral and written communications. It is true that they must reflect careful thought and precise expression, since they will be interpreted critically, and since they stand and must be judged in isolation. Thus they must be self-contained in meaning, depending on content not on context. But the problem of true-false item writing is no different from the problem of writing for any other purpose of communication. Those who have difficulty in writing good true-false test items probably have trouble expressing themselves clearly and accurately in other situations also.

Two Requirements for Writing Good True-False Test Items

The first requirement is mastery of the subject to be tested, which in the case of teachers always implies mastery of the language in which knowledge of the subject is expressed. In most subjects this mastery grows slowly over many years. It can seldom be acquired in a single course, and never in a course on test construction. But to the degree that it is lacking the tests produced are likely to be deficient, and no amount of special training in test construction can make up for the deficiency. Many of the shortcomings of teacher's tests, including their true-false tests, are due to their inadequate command of the knowledge they are trying to test.

But other shortcomings are due to lack of knowledge of special techniques of item writing and test construction. There are tricks-of-the-trade, knowledge of which will help any teacher to make better tests. To point out some of these that are useful in writing true-false test items is the principal mission of this article.

The Process of Writing True-False Test Items

A good true-false test item is one that is both acceptable and effective. It will be acceptable as a measure of achievement, to one who believes the central purpose of formal education is to gain command of useful verbal knowledge, if it is significant and supportable. To be significant it must test the student's command of some important element of useful knowledge that is not common knowledge. To be supportable it must be unquestionably true or false in the opinion of a knowledgeable expert. It will be

effective if those who lack an adequate grasp of that element of knowledge find a wrong answer attractive.

There are thus four essentials to writing a good true-false test item.

1. Choose a significant idea.
2. Devise a problem which will require understanding or application of the idea.
3. Word the statement of the problem so that those who lack understanding of the element being tested will be attracted to a wrong answer.
4. Review the statement critically to make sure that any who do understand the point being tested ought to answer it correctly.

The Problem of Significance

A test item is significant if it deals with an element of knowledge that is part of a structure of related concepts, ideas or events, and that is likely to be useful on future occasions. Significant items in a test are those that deal with the ends rather than the means of instruction. When a messenger knocks, it is the message he bears that is usually significant, not whether he knocked three times or two. About every item in a good test it should be possible to give an affirmative answer to the question, "Does this item test an element of knowledge that is really worth knowing?"

It is good for a test constructor to make each item he writes as significant as he can. But if he is writing objective test items he should not expect each of them to appear to have tremendous individual significance. There are, after all, a great many of them in the test. Each one involves directly only one element in a vast and complex structure of knowledge. The substantial significance of a test score rests on the summation of many lesser significances.

There is a second reason why the items of even the best true-false tests may not appear to be highly significant individually. It is because of the typical inverse relation between an item's apparent significance, and the definiteness of its truth or falsity. The requirement that each true-false test item be unequivocally true or false rules out the use of broad generalizations that might appear to be highly significant, but whose truth has not been and

perhaps cannot be definitely established. Thus the loss to the test constructor is only apparent, not real. The generalization may deal with a very important problem, but if its truth is indeterminable it can not give a significant indication of achievement. Regardless of what form of test item is used, essay or objective, one can not obtain definite assessments of competence by asking questions that have indefinite answers.

Faced with the difficult problem of finding true propositions of high apparent significance to use as the basis for test items, it is reassuring to recall that, as Howard's (Howard, 1943) study showed, there is a substantial correlation between the indications of achievement given by items of less, and those of more, apparent significance. This means that the degree of apparent significance probably is not a crucial factor in the validity of the measures of achievement obtained. The difference between good and poor command of knowledge in an area of study shows up about as clearly on the less significant items as it does on the more significant ones.

It remains true, however, that the acceptability of a test depends substantially on the apparent significance of the items composing it. The test constructor is well advised to try to use as many highly significant items in his test as he can succeed in developing.

Testing Command of Knowledge

To test a person's command of an idea or element of knowledge is to test his understanding of it. A student who can recognize the words in which an idea has been expressed but who cannot recognize the same idea when it is expressed in different words does not have command of it. Or, if he knows the idea only as an isolated fact, without seeing how it is related to other ideas he has no command of it. Knowledge one has command of is not a miscellaneous collection of separate elements. It is an integrated structure. Knowledge one has command of is knowledge one can use to make decisions, draw logical inferences, or solve problems. It is usable knowledge.

Consider how one might test a student's command of Archimedes Principle. It should not be done by offering him the usual expression of the principle as a true statement, or some slight

alteration of it as a false statement, as has been done in items 1 and 2 below.

1. A body immersed in a fluid is buoyed up by a force equal to the weight of the fluid displaced. (T)
2. A body immersed in a fluid is buoyed up by a force equal to half of the weight of the fluid displaced. (F)

Instead the student might be asked to recognize the principle in some alternative statement of it, such as in items 3 and 4 below.

3. If an object having a certain volume is surrounded by a liquid or gas, the upward force on it equals the weight of that volume of the liquid or gas. (T)
4. The upward force on an object surrounded by a liquid or gas is equal to the surface area of the object multiplied by the pressure of the liquid or gas surrounding it. (F)

Or the student might be required to apply the principle in specific situations such as those described in items 5 and 6 below.

5. The buoyant force on a one centimeter cube of aluminum is exactly the same as that on a one centimeter cube of iron when both are immersed in water. (T)
6. If an unsoluble object is immersed successively in several fluids of different density, the buoyant force upon it in each case will vary inversely with the density of the fluids. (F)

Sometimes the use of an unconventional example can serve to test understanding of a concept.

7. Distilled water is soft water. (T)

It is a popular misconception that true-false test items are limited to testing for simple factual recall. On the contrary, complex and difficult problems can be presented quite effectively in this form.

8. The next term in the series 3, 4, 7, 11, 18, is 29. (T)
9. If the sides of a quadrilateral having two adjacent right angles are consecutive whole numbers, and if the shortest side is one of the two parallel sides, then the area of the trapezoid is 18 square units. (T)

True-false test items can also be used to test command of knowledge in fields of study such as history and literature. Here the facts are likely to have specific rather than general significance but they are the elements of historical and literary knowledge not isolated details, and ought not to be learned as such. They are parts

of a structure of knowledge, less universal than scientific knowledge perhaps, but none the less a structure. A student who has command of an element of literary or historical knowledge will understand the ideas involved in it, will be able to draw inferences from them, and will know their relations to other ideas.

Consider, as an illustration, the episode in American history known as the Battle of Trenton, described in the passage below. A student's command of this segment of historical knowledge can be tested by using true-false items such as those numbered 10 to 15 below.

The Battle of Trenton¹

The attack on the Hessian troops stationed at Trenton was made at dawn on December 26, 1776. During the previous night George Washington and 2,500 of his troops had crossed the Delaware River from Pennsylvania through floating ice. They landed in New Jersey about nine miles above Trenton. Approaching the town by two roads, the American army surprised the Hessian outposts and then rushed upon the main body before it could form effectively. The charge of the American troops and their fire of the artillery and muskets completely disconcerted the enemy. A few hundred escaped but the majority (over 900) were surrounded and forced to surrender.

10. Before the Battle of Trenton the American Army crossed the Delaware River from west to east. (T)
11. The Battle of Trenton took place during the second year of the Revolutionary War. (T)
12. The American troops outnumbered the British troops in the Battle of Trenton. (T)
13. Surprise was a major factor in the outcome of the Battle of Trenton. (T)
14. The American army suffered few casualties in the Battle of Trenton. (T)
15. The Battle of Trenton was George Washington's first major success in the Revolutionary War. (T)

Note that none of these items deals simply and directly with any of the specific statements in the description of the battle. All of

¹ Adapted from the Encyclopedia Britannica Vol. 22 page 218, 1968.

them require inferences from these facts, or understanding of the relation between this event and other aspects of the war. By so doing, they seek to test command of the knowledge.

The Problem of Definiteness.

A good true-false test item is definitely true or false in the eyes of qualified experts. This means that the point involved must be well-established truth. It means that the point must be expressed clearly. It means that concise statements of the point should be sought, since conciseness usually contributes more to clarity than does complexity.

Three things can be done beyond those already implied to help solve the problem of definiteness in writing true-false test items. One is to write, or at least to think of the items in pairs, one true and one false such as items 16 to 19 below.

16. An eclipse of the moon can only occur when the moon is full. (T)
17. An eclipse of the moon can only occur when the moon is new. (F)
18. The average farm in the United States was larger in 1960 than it was in 1910. (T)
19. The average farm in the United States was smaller in 1960 than it was in 1910. (F)

This procedure helps to clarify what the item is testing and to indicate whether or not it is worth testing. It encourages conciseness of expression. It helps avoid items like 20 and 21 below which have no plausible alternatives and which therefore would make poor true-false test items.

20. Insurance agencies may be either specialized or general. (T)
21. Camping has a good past, a better present and an almost unlimited future. (T)

Of course only one member of any pair is used in the same test. The other is sometimes sufficiently different to be usable in a second test, or in a different form of the test.

A second thing that can be done to help solve the problem of definiteness is to write statements which call for comparison between two specified alternatives as in items 22 and 23 below.

22. The time from moon rise to moon set is generally longer than the time from sunrise to sunset. (T)

23. The beneficial effect of a guessing correction, if any, is more psychological than statistical. (T)

Such internal comparison focuses attention clearly on the essential question in the item. Of even greater help is the fact that it avoids the necessity of using arbitrary standards in judging truth or falsity and the resulting possibility that the examiner's standards might differ significantly from those of the examinee.

The third thing which helps to avoid indefiniteness in true-false test items is careful review of the items after they have been written. This can be of value even if done by the author of the items himself, after several days have passed, and after the context in which the items were written has been forgotten. It can be of even greater value if done independently by a competent colleague. Independent review is not likely to supply quality to true-false test items that are grossly lacking in it, but it can help to avoid errors and ambiguities in communication that sometimes result from singularity in point of view or mode of expression.

Making Wrong Answers Attractive.

The job of a test item is to discriminate between those who have and those who lack command of some element of knowledge. Those who have the command should be able to answer the question correctly without difficulty. Those who lack it should find the wrong answers attractive. To make them so is one of the arts of item writing. Here are some of the ways in which it can be done.

- A. Use more false than true statements in the test.

When in doubt, students seem more inclined to accept than to challenge propositions presented in a true-false test. The experimental evidence for this inclination is impressive. Of course if they come to expect more false items than true, some of the value of this technique is lost. But the imbalance is not easy to discover, and cannot be counted on confidently in the future even if discovered. So it continues to work quite well even in classes which are aware of it as a possibility.

- B. Word the item so that superficial logic suggests a wrong answer.

24. A rubber ball weighing 100 grams is floating on the surface of a pool of water exactly half submerged. An additional

downward force of 50 grams would be required to submerge it completely. (F)

The ball is half submerged and weighs 100 grams, which gives one half of 100 considerable plausibility on a superficial basis. The true case is, of course, that if its weight of 100 grams submerges only half of it, another 100 grams would be required to submerge all of it. Superficial logic also would make the incorrect answers to these questions seem plausible.

25. Since students show a wide range of individual differences, the ideal measurement situation would be achieved if each student could take a different test specially designed to test him. (F)

26. The output voltage of a transformer is determined in part by the number of turns on the input coil. (T)

27. A transformer that will increase the voltage of an alternating current can also be used to increase the voltage of a direct current. (F)

C. Make the wrong answer consistent with a popular misconception or a popular belief irrelevant to the question.

28. The effectiveness of tests as tools for measuring achievement is lowered by the apprehension students feel for them. (F)

Many students do experience test anxiety, but for most of them it facilitates rather than impedes maximum performance.

29. An achievement test should include enough items to keep every student busy during the entire test period. (F)

Keeping students busy at worthy educational tasks is usually commendable, but in this case it would make rate of work count too heavily, in most cases, as a determinant of the test score.

D. Use specific determiners in reverse to confound test wiseness.

In true-false test items extreme words like *always* or *never* tend to be used mainly in false statements by unwary item writers, whereas more moderate words like *some*, *often*, or *generally* tend to be used mainly in true statements. When they are so used they qualify as "specific determiners" which help testwise but uninformed examinees to answer true-false questions correctly. But some *always* or *never* statements are true and some *often* or *generally* statements are false. Thus these specific determiners can

be used to attract the student who is merely testwise to a wrong answer.

E. Use phrases in false statements that give them the "ring of truth".

30. The use of better achievement tests will, in itself, contribute little or nothing to better achievement. (F)

The phrases "in itself" and "little or nothing" impart a tone of sincerity and rightness to the statement than conceals its falseness from the uninformed.

31. To insure comprehensive measurement of each aspect of achievement, different kinds of items must be specifically written, in due proportions, to test each different mental process the course is intended to develop. (F)

There is superficial logic to this statement like those illustrated under B above. But it also displays the elaborate statement and careful qualifications that testwise individuals associate mainly with true statements.

Is a teacher playing fair with his students if he sets out deliberately to make it easy for some of them to give wrong answers to his test items? If he wants to measure achievement validly, that is to distinguish correctly between those who have and those who lack command of a particular element of knowledge, it is the only way he can play fair. The only reason a test constructor sets out to make wrong answers attractive to those who lack command of the knowledge is so that correct answers will truly indicate the achievement they are supposed to indicate.

Conclusions

This article has attempted to set forth some reasons why true-false test items should be used in measuring educational achievement, and some means by which they can be used effectively. In the author's teaching experience during the last decade, they have proved satisfactory to him and to his students. They have shown up well in test analysis. It seems likely that some teachers who have been touted off true-false tests could serve themselves and their students well by taking another close look at them.

REFERENCE

Howard, Frederick T. *Complexity of mental processes in science testing*. Contributions to Education No. 879, Teachers College, Columbia University, New York. 1943.

A NOTE ON GAYLORD'S "ESTIMATING TEST RELIABILITY FROM THE ITEM-TEST CORRELATIONS"

JOHN BOWERS
University of Illinois

GAYLORD (1969) demonstrated the algebraic inconsistency of Guilford's (1956) reliability formula,

$$r_{tt} = \frac{n\bar{r}_{it}^2}{1 + (n-1)\bar{r}_{it}^2}, \quad (1)$$

based upon Richardson's (1936) relationship,

$$\bar{r}_{ij} = \bar{r}_{it}^2, \quad (2)$$

where \bar{r}_{ij} = the average item intercorrelation and \bar{r}_{it}^2 = the square of the average item-test correlation.

It may be instructive to examine this inconsistency. When item variances are assumed equal, the sum of all elements in the item variance-covariance matrix is,

$$\sum_i \sum_j r_{ij} \sigma_i^2 = \sigma_t^2, \quad (3)$$

where

$i = 1, \dots, n$ rows,

$j = 1, \dots, n$ columns,

$r_{ij} = 1$ when $i = j$,

σ_t^2 = total score variance,

σ_i^2 = the common item variance.

The sum of all elements in the i -th row is,

$$\sum_j r_{ij} \sigma_i^2 = r_{it} \sigma_i \sigma_t, \quad (4)$$

and from (3) and (4), when σ_i and σ_t are cancelled,

$$\sum_i r_{it} = \sqrt{\sum_i \sum_j r_{ij}}. \quad (5)$$

When (5) is squared and averaged over n items,

$$\bar{r}_{ii}^2 = \bar{r}_{ij}. \quad (6)$$

This is Richardson's (1936) identity, and is the average of the elements in the item intercorrelation matrix including ones in the diagonal.

If \bar{r}_{ij}' is defined as the average of the off-diagonal elements in the item intercorrelation matrix, then as Gaylord has shown,

$$\bar{r}_{ij}' = \frac{n\bar{r}_{ii}^2 - 1}{n - 1} = \frac{n\bar{r}_{ij} - 1}{n - 1}. \quad (7)$$

When the reliability of each item in a test is estimated by its average correlation with the remaining $n - 1$ items in a test (Kuder and Richardson, 1937), the Kuder-Richardson reliability estimate is

$$r_{tt} = \frac{\sigma_t^2 - n\sigma_i^2(1 - \bar{r}_{ij}')}{\sigma_t^2}, \quad (8)$$

which can be reduced to,

$$r_{tt} = \frac{n\bar{r}_{ij}'}{1 + (n - 1)\bar{r}_{ij}'} \quad (9)$$

Guilford's (1956) formula is erroneous, since $\bar{r}_{ii}^2 = \bar{r}_{ij}$ rather than \bar{r}_{ij}' . But, as (7) indicates, the error diminishes as n becomes large.

Gaylord's interesting expression (7) is, in this notation,

$$r_{tt} = \frac{\bar{r}_{ij}'}{\bar{r}_{ij}}. \quad (10)$$

Thus, KR-20, under the assumption of equal item variances, is the ratio of two averages calculated from the item intercorrelation matrix. When item means are also equivalent, expression (10) is the KR-21 reliability estimate.

The true score variance, σ_ω^2 , is defined as the product of the observed variance and the reliability. From (3),

$$\sigma_t^2 = n^2\sigma_i^2\bar{r}_{ij}, \quad (11)$$

so, multiplying (10) by (11),

$$\sigma_\omega^2 = n^2\sigma_i^2\bar{r}_{ij}'. \quad (12)$$

The definition of reliability as the ratio of true score to observed

score variance holds since (12) divided by (11) equals (10). The error variance is the difference between (11) and (12),

$$\sigma_e^2 = n^2 \sigma_i^2 (\bar{r}_{ii} - \bar{r}_{ii}') \quad (13)$$

or

$$\sigma_e^2 = n^2 \sigma_i^2 \left(\frac{1 - \bar{r}_{ii}'}{n} \right). \quad (14)$$

which shows, as is well known, that Kuder-Richardson reliabilities depend altogether on item intercorrelations.

REFERENCES

- Gaylord, R. H. Estimating test reliability from the item-test correlations. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 303-304.
- Guilford, J. P. *Fundamental statistics in psychology and education*. (3rd ed.) New York: McGraw-Hill, 1956.
- Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, 2, 151-160.
- Richardson, M. W. Notes on the rationale of item analysis. *Psychometrika*, 1936, 1, 69-76.

THE EFFECTS OF FOREWARNING AND PRETESTING ON ATTITUDE CHANGE

GLORIA COWAN

Wayne State University

S. S. KOMORITA

Indiana University

McGUIRE (1968) has commented that, "... today's artifact may be tomorrow's independent variable." An instance of the validity of this hypothesis is illustrated in the field of persuasive communication where the subject's suspiciousness of the experimenter's intent to persuade has become a substantive issue in its own right. In earlier years, the problem of suspicion and intent of the experimenter was studied with a focus on the social psychology of the experiment. Orne (1962) suggested that the subject formulates his own hypotheses about the nature and purpose of the experiment, and therefore, the subject's behavior—derived from the demand characteristics of the experiment—may then be a response to his own hypotheses. Silverman (1968), for example, found more acquiescence to a persuasive message when it was presented in the context of a psychological experiment than when it was not.

In recent years, however, two opposing hypotheses, aside from the methodological implications of the problem, have been proposed: Hovland and Festinger's classical view that awareness and warning evoke resistance and decrease the persuasive impact of a communication and McGuire's hypothesis (1968) that suspiciousness of intent may actually enhance the impact of the message by increasing message reception.

The main purpose of this experiment was to explore the rela-

tionship between the subject's awareness of the intent of the experimenter, Orne's "demand characteristics," and attitude change. The extent to which the subject's perception of how he is expected to respond to a communication is related to his actual response in an attitude change context has both methodological and substantive implications. The substantive issue of forewarning may be tested by experimentally forcing the subject to indicate his perception of the intent of the experimenter *before* he indicates attitude change (forewarning) vs. after he indicates attitude change (no forewarning.)

An essential mediating step between general suspiciousness of the intent of the experimenter and attitude change is the direction and degree of change the subject thinks the experimenter expects him to show. If all subjects' evaluations of the intent of the experimenter are the same, a unidirectional and constant effect of suspiciousness on attitude change would be predicted, and forewarning should increase (or decrease) the impact of the communication—depending on what theoretical position, Hovland or McGuire, is taken. However, if the mean amount of attitude change of a forewarned group does not differ significantly from that of a group not forewarned, and the particular hypotheses vary with the subject, forewarning should increase the correlation between the subject's hypothesis and his experimental behavior.

If we look hypothetically at the possible kinds of hypotheses available to the *S*, there are three possibilities. The *S* may hypothesize that: (a) he is expected to change in a favorable direction, (b) he should resist influence and should show that he is not so easily persuaded, and (c) the intent of the *E* is to show the fallacy of a one-sided communication, thus suggesting the opposite or a boomerang effect. Rather than to assume similar demand characteristics on the part of different subjects, the hypotheses or suspicions may be directly measured.

A secondary purpose of the study was to determine if pretesting sensitizes the subject to the experimental treatment by providing additional demand characteristics than those available in the "after-only" design. Pretesting can also be viewed as a suspicion arouser, decreasing attitude change, or an enhancer of attitude change by increasing the clarity of the demand characteristics. Although there is little evidence to indicate that the pre-

test sensitizes subjects to the experimental treatment (e.g., Lana, 1959), the pretest, if anything, tends to reduce the effect of the experimental treatment (Hovland, Lumsdaine, and Sheffield, 1949). Accordingly, it is plausible that the effects of the "pre-post" design, as compared to the "after-only" design, may interact with the demand characteristics of the experiment. Initial attitude toward the issue provides, at the same time, another link in the suspiciousness change chain. Although the subject's initial attitude toward the issue has been investigated via discrepancy between position urged in the message and *S*'s initial position and attitude change, it is plausible that the perception of the intent of the message, and thus, the direction of suspiciousness itself, is related to the subject's initial position. Again, pretesting, like forewarning, may fail to exert a constant effect (positive or negative) on attitude change, but instead may increase the correlation between the subject's own hypothesis and his experimental behavior.

Procedure

Seventy-five students in Introductory Psychology classes were given a favorable written communication about advertising.¹ Half had been pretested two weeks earlier in their classes with the advertising issue embedded in nine other issues. The other half were not given a pretest. Both the pretest and posttest consisted of five evaluative semantic differential scales.

All subjects were asked to read and to evaluate an article on advertising appearing in the *Saturday Review* by Charles Horton, an expert on advertising. Before reading the article, the subjects were told that they would be asked to evaluate the readability of the article and the point of view of the author. Immediately after reading the communication, half the subjects took the posttest and half took the awareness measure. Those subjects who responded to the posttest first then responded to the awareness measure, and those who took the awareness measure first then responded to the posttest. Subjects were run in two large groups.

¹ "Advertising" was chosen as an issue because pretesting on various issues indicated that the mean on the advertising issue was 20.7, close to the scale neutral point of 20, and the standard deviation of 8.2 allowed for variability in initial attitude.

The instruction given orally for the awareness measure after the posttest were:

We would like you to fill out one more questionnaire, but the instructions for this questionnaire are different from the ones you were given on the previous questionnaire. What we would like you to do this time is to show how you think we expected you to respond to the article. Every psychological experiment starts with a hypothesis and the experimenter is making predictions about how you will respond. What we want you to do is to show us what you think we are predicting.

You can tell us your ideas about what the experiment is about by filling out the same sheet you just completed. This time answer it in the way in which you think we expected you to respond. By filling out this questionnaire, you are telling us what you think is the purpose of this experiment. Does everyone understand what he is asked to do? Remember, this time you are showing us what you think the experiment is about, rather than how you feel about advertising.

This may or may not be the same as your previous responses.

The instructions for responding to the awareness measure for those subjects who had not yet taken the posttest were essentially the same, differing only in those aspects referring to the posttest.

The awareness instrument was the identical semantic differential used as pre and post tests. The semantic differential awareness measure forces the subject to indicate quantitatively the direction and amount of change he thinks he is expected to show, and does not permit him to respond ambiguously about the intent of the experimenter. It is also possible to determine the difference between his hypothesis and his experimental behavior using this instrument.

Giving the awareness measure before the posttest but not before the presentation of the communication eliminates the possibility of the subject's practicing or elaborating his defense.

For all groups, the correlation coefficients between awareness and posttest measures were obtained. For the pretested groups, the correlation coefficients between pre and post measure and between pre and awareness measures were also obtained. The design of the study was a 2×2 factorial with forewarning (pre-

resentation of the awareness measure before vs. after the posttest measure) and pretesting vs. no pretesting as the two variables.

Results

Mean differences. Table 1 presents the means of the three response measures: posttest scores, awareness scores, and change scores as a function of forewarning and pretesting. The analyses of variance of the data in Table 1 indicated that neither the main effects of forewarning or pretesting nor the interaction between the two variables was significant at the .05 level. Thus, no significant *directional* effect of forewarning or pretesting was found on subjects' responses to the communication or his awareness of how he was supposed to respond.

Correlational analyses. The correlation coefficients between pretest, awareness, and posttest measures are shown in Table 2. It can be seen that the correlations differ significantly as a function of forewarning and pretesting. A significant relationship between awareness and posttest scores was obtained only in the pretested groups given the awareness measure prior to the posttest. Since the sample sizes were quite small, it was decided to replicate the study for the pretest groups, particularly so that pretest level could be assessed. Table 2 also shows the data for the replication (Sample 2). It can be seen that a similar pattern of correlations was obtained for Sample 2, indicating that these differences are quite reliable.

For the combined pretested groups, the .60 correlation between awareness and posttest for the forewarned group is significantly greater than the .01 correlation for the non-forewarned group

TABLE 1
Mean Posttest, Awareness, and Change Scores as a Function of Order and Pretesting

Means	Treatment Groups			
	Pretested, Posttest first (<i>n</i> = 19)	Pretested, Awareness first (<i>n</i> = 18)	Not Pretested Posttest first (<i>n</i> = 18)	Not Pretested Awareness first (<i>n</i> = 20)
Posttest	23.36	25.83	23.55	24.80
Awareness	27.36	30.00	26.11	27.40
Change (post-pre)	2.63	5.11	—	—

TABLE 2
Correlations between Pretest, Posttest, and Awareness

	Posttest Awareness	Pre-post	Pretest Awareness
A. Pretested Groups			
1. Posttest first			
Combined ($n = 57$)	.01	.69**	.33**
Sample 1 ($n = 19$)	.18	.70**	.41*
Sample 2 ($n = 38$)	.07	.69**	.29*
2. Awareness first			
Combined ($n = 42$)	.60**	.09	.08
Sample 1 ($n = 18$)	.57*	.24	.07
Sample 2 ($n = 24$)	.63**	.05	.09
B. Post-only groups			
1. Posttest first			
Sample 1 ($n = 18$)	.07	—	—
2. Awareness first			
Sample 1 ($n = 20$)	.26	—	—

* $p < .05$.** $p < .01$.

(one-tailed tests)

($z = 3.25$, $p < .01$), but is not significantly greater than the .26 correlation in the forewarned but not pretested group ($z = 1.35$). Column 2 of Table 2 also shows that there is a high correlation between pre and post test for the non-forewarned group (.69), but very little relationship between pre and post test for the forewarned group (.09). The difference between the two correlations is significant at the .01 level ($z = 3.73$).

There is a significant relationship between pretest and awareness scores, as seen in column 3 of the correlational table, for the non-forewarned group (.33) but not for the forewarned group (.08); however, the difference between the two correlations is not significant ($z = 1.23$). Thus, although forewarning and pretesting do not affect mean awareness or posttest scores, they do appear to affect the relationship between these measures.

A supplementary analysis was conducted to test the interaction between pretest level and awareness; i.e., to determine whether those subjects both initially favorable and aware were most likely to obtain high posttest scores. Accordingly, all subjects who had been pretested were divided into high and low initial attitude groups and high and low awareness groups (median splits). The analysis of variance showed a significant main effect of initial attitude with subjects with higher initial pretest scores, as might be

expected, also scoring higher on the posttest ($F = 13.20, p < .01$). There was also a tendency, of borderline significance ($F = 3.86, p < .06$), for more aware subjects to have higher posttest scores than less aware subjects, but the interaction between awareness and pretest level was not significant.

Discussion

The results of this study indicate a significant relationship between the subject's awareness of the intent of the experimenter and his responses to a communication if and only if he is both forewarned and pretested. When the subject has been pretested and is forced to formulate and make explicit his suspicions prior to posttesting, his posttest attitude toward the issue is significantly related to his specific suspicions. Apparently, asking the subject to give his hypothesis first allows him to use this new cue, his experimental hypothesis, to direct his experimental behavior. On the other hand, since the mean posttest scores do not vary as a function of experimental conditions, suspiciousness or pretesting does not seem to produce a directional effect on attitude change, neither resistance to or facilitation of persuasion. Under conditions that make suspiciousness salient, the particular suspicions or expectations of how he should respond are related to the subject's actual responses. Thus, suspiciousness may be operating in an experiment, *when aroused*, but may be masked if the experimental demands, although salient, are sufficiently ambiguous to allow variability in how subjects think they are expected to respond.

The ambiguity of previous findings on warning of persuasive intent (McGuire, 1968) can be accounted for in these terms in that the nature of the forewarning by the experimenter cannot be assumed to have a direct relationship to the way the subject interprets the forewarning. Hastorf and Piper (1951), for example, found no resistance to suggestion produced by explicitly reminding the subjects that they had answered a pretest and should give similar answers after receiving some normative feedback. It is possible, therefore, that some subjects in the Hastorf and Piper study deduced that they were expected to resist the *experimenter's* instructions not to change. Even an explicit statement of intent to persuade may not lead to similar deductions by subjects that the experimenter expects them to heed the forewarning.

In the present study, the pattern of relationships between pretest, posttest, and awareness for the treatment group for whom suspicion should be greatest (pretested and forewarned) also suggests that forewarning greatly attenuates the relationship between initial attitude toward the issue and attitude after the presentation of the persuasive communication. At the same time, it is clear that the subject's hypotheses regarding the intent of the experimenter, when forewarned, are independent of his initial favorability toward the issue.

In the nonforewarned group, the relationship between pre and posttest is strong, but there is no relationship between awareness of intent and posttest. As the subject has not been forced to make explicit his hypotheses before responding to the communication, his response to the communication does not seem to be influenced by his specific hypotheses, although the degree of favorability toward the issue he thinks he is expected to show is moderately related to his initial favorability toward the issue. Thus, initial attitude toward the issue does not appear to provide a link in understanding the relationship between the subject's specific suspicions and his response to the communication.

Pretesting alone and forewarning alone are not sufficient suspicion arousers to affect either the direction of response to the communication or the relationship of response to the communication with the suspicion of the subject; however, both forewarning and pretesting suggest that the subject will respond to the communication in a way consistent with how he thinks he is expected to respond.

From a substantive perspective, these findings support the view that salience of intent to persuade operates, but selectively. The lack of clear support for forewarning as a facilitator vs. forewarning producing resistance may be due, in part, to the possibility that forewarning has variable effects on the particular suspicions the subject subsequently formulates; consequently, the facilitating effects of the subject's particular suspicions may not be obvious if the specific suspicions are not assessed.

From a methodological perspective, our findings do not support Orne's objection to the use of post-experimental inquiries as measures of the subject's hypotheses about the purpose of the experiment on the basis that the hypotheses are influenced by the

subjects preceding experimental behavior. Moreover, it does not seem feasible to assess the hypotheses before the subject responds to the communication because his response to the communication will be affected by the nature of his hypothesis. This conclusion, of course, is qualified by the nature of the awareness measure used in this experiment.

The concern about the sensitizing effects of pretesting, *per se*, does not seem to be warranted since there is no significant difference between posttest scores or awareness scores of those subjects pretested vs. those subjects who were not pretested. Pretesting, itself, does not seem to provide additional demand cues to the subject. The absence of a pretesting sensitization effect should be qualified in terms of the precautions taken in this study to separate the pretest phase from the posttest phase of measurement; i.e., embedding pretest with other measures, spacing pretest and treatment, using different experimenters for pretest and treatment. The major methodological implications of this study are that all the concern over sensitization may have been overemphasized and that to the extent that these findings may be generalized, it does not seem necessary to use Solomon's four group design to control for pretest sensitization nor an awareness control group.

REFERENCES

- Hastorf, A. H. and Piper, G. N. A note on the effect of explicit instructions on prestige suggestion. *Journal of Social Psychology*, 1951, 33, 289-293.
- Hovland, C. I., Lumsdaine, A. A., and Sheffield, F. D. *Experiments on mass communication*. Princeton, New Jersey: Princeton University Press, 1949.
- Lana, R. E. Pretest-treatment interaction effects in attitudinal studies. *Psychological Bulletin*, 1959, 56, 293-300.
- McGuire, W. J. Suspiciousness of experimenters intent as an artifact in research. In Rosenthal, R. and Rosnow, R. (Eds.) *Artifact in Social Research*. New York: Academic Press, 1968.
- Orne, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 1962, 17, 776-783.
- Silverman, I. Role-related behavior of subjects in laboratory studies of attitude change. *Journal of Personality and Social Psychology*, 1968, 8, 343-348.

A COMPARATIVE STUDY OF FIVE METHODS OF ASSESSING SELF-ESTEEM, DOMINANCE, AND DOGMATISM¹

DAVID L. HAMILTON

Yale University

THE past two decades have seen a rapid growth in the number of personality scales available for use in research concerned with individual differences in experimentally-studied behaviors. Consequently, one finds several scales with the same trait name or which appear to be measuring the same or a similar construct. In comparing results of experiments which have employed these different instruments, one wonders whether the same personality attribute is indeed being assessed by the different methods. Without evidence to this effect, generalization across studies is at least risky, at most, unwarranted.

The present study was undertaken to examine the construct validity of several commonly-used personality measures. Campbell and Fiske (1959) have pointed out that correlational evidence for the construct validity of a psychological test requires demonstration of both convergent validity and discriminant validity. They also point out that two tests may correlate highly because they have method, as well as trait, variance in common. From these considerations it follows that correlational evidence for construct validity requires that each of at least two traits be measured by at least two methods. The correlations may be presented in what Campbell and

¹ The data for this research were collected while the author was at the University of Illinois. The study was supported by a grant from the United States Public Health Service, National Institutes of Health, No. M-4460 (Ivan D. Steiner, Principal Investigator). The author expresses appreciation to J. Richard Hackman for his comments on an earlier version of the manuscript.

Fiske called a multitrait-multimethod matrix, which may then be examined for the convergent and discriminant validity of the various measures.

A variety of methods have been advocated for the assessment of personality attributes. Probably the most commonly-used approach to personality scale construction has been the method of empirical keying, in which items gain meaning only through their demonstrated ability to differentiate between groups known to differ on some external criterion (Meehl, 1945). Although inventories developed by this methodology have found wide use, the interscale correlations have frequently been high enough to question the discriminant validity of scales with quite diverse trait names. Alternatively, some writers (e.g., Loevinger, 1957) have argued that test construction should have its roots in a conceptual description of the construct of interest, so that items in the measuring instrument will reflect the properties of the attribute being assessed. Peterson (1965) has advocated an even more direct approach in which the subject simply rates himself on the attribute of interest. He described a study by Wetzel (1963) in which simple self-ratings on adjustment and introversion-extraversion correlated highly with ratings on these traits made by parents and peers, while intertrait correlations were small. Peterson argued that the evidence for the convergent and discriminant validity of lengthy inventories is no more compelling than that reported by Wetzel for the simple self-ratings, at least for these two attributes. As opposed to these self-report methods, the use of peer ratings has been proposed as an assessment device which avoids many of the problems inherent in self-descriptive techniques (Smith, 1967).

This study examined the convergent and discriminant validity of alternative methods of measuring self-esteem, dominance, and dogmatism. These attributes were selected because (a) they seem to be of primary interest to psychologists (as reflected in the literature) and (b) several methods of assessing these traits have been proposed. Five methods of measuring these attributes were compared: empirically-derived true-false inventory scales, conceptually-based self-descriptive questionnaires, conceptually-based checklist measures, simple self-ratings, and peer ratings.

Method

Subjects

Subjects in the experiment were 70 male undergraduate students at the University of Illinois. All subjects belonged to one of two fraternities, 36 subjects belonging to one fraternity and 34 to the other. Subjects were not paid individually, but each fraternity was paid for its participation.

Procedure

Subjects were met in the fraternity house and all subjects in a given fraternity were tested at the same time. A packet of test materials was given to each subject, the packet consisting of several self-rating questionnaires and a form for making peer nominations. These instruments are described below.

Measures

Method I consisted of measures of the various traits by true-false, empirically derived inventory scales. Specifically, items from six scales of the California Psychological Inventory (CPI) were given to the subjects. Among them were the Dominance and Flexibility scales and two scales which, on the basis of descriptions in the CPI Manual (Gough, 1957), were considered relevant to the more global concept of self-esteem, the Social Presence and Self-Acceptance scales. Scores on the latter two scales were summed to provide an index of self-esteem.

Method II consisted of measures in which each item requires the subject to rate himself on an intensity continuum. Only two instruments were of interest here: the Dogmatism scale (Rokeach, 1960) and the Janis-Field Feelings of Inadequacy scale (Janis and Field, 1959), which has frequently been used as a measure of self-esteem in attitude change research. The present writer knows of no previous investigation of this scale's validity, in spite of its wide use in research on persuasibility. The scale was scored such that high scores reflected high self-esteem (i.e., few "feelings of inadequacy"). It should be noted that the Dogmatism and Feelings of Inadequacy scales differ from the CPI measures of Method I in (at least) two important ways. First, these scales ask the subject to indicate the degree to which the statements are descriptive of his feelings and

attitudes, instead of employing dichotomous true-false response categories. Second, the item content of these Method II measures was designed to reflect the various properties of the construct being assessed, whereas item content is given less consideration in tests developed by a criterion-group methodology.

Method III consisted of measures derived from the Leary Interpersonal Checklist (ICL) (Leary, 1957). The two main dimensions in the ICL are called Dominance-Submission and Love-Hate. The Checklist was scored for the 16 categories of interpersonal behavior represented by the item domain, and Dominance and Love scores were determined for each subject using the formulas suggested by LaForge (1963). The former scores provided a checklist measure of dominance. A measure of self-esteem was then determined which, although not a checklist measure in the narrower sense, was derived from the Interpersonal Checklist. Subjects completed the ICL twice, once checking those items characteristic of themselves, and later, on a separate ICL form, checking those items characteristic of the way they would ideally like to be. A small discrepancy between self and ideal-self concepts has often been considered indicative of a high degree of self-satisfaction, while a large discrepancy suggests a weak self-concept or low self-esteem. Dominance and Love scores were determined for each subject for both forms, thereby determining the location of each subject's Self and Ideal-Self concepts within the circular framework on the ICL. The geometric distance between the Self and Ideal-Self (the Self-Ideal Distance) was then computed (LaForge, Leary, Nabolssek, Coffey, and Freedman, 1954). This distance was then taken as an index of the extent of discrepancy between Self and Ideal-Self, small distances being considered indicative of high self-esteem.

Method IV consisted of simple self-ratings. Subjects rated themselves on three seven-point scales similar in format to those used by Wetzel (1963). The scales were labelled "Dominant-Submissive," "Closed-vs. Open-Minded," and "High vs. Low Self-Esteem."

Method V consisted of peer nominations. The procedures employed were similar to those described by Norman (1963). The two poles of each of 11 dimensions (e.g., "calm-anxious") were listed separately, making 22 attributes on which peer nominations were made. For each attribute, each subject was to name the four members of his fraternity group (limited to those participating in the

study) most characterized by the trait. Peer nomination scores were then determined as follows: the number of times a person was nominated for the second pole of a dimension was subtracted from the number of times he was nominated for the first pole, and a constant of 35 was added to eliminate negative values. Two of the dimensions—"Has High Self-Esteem-Feels Inferior" and "Self-Confident-Lacks Confidence"—were considered relevant to self-esteem and these scores were summed for a measure of Self-Esteem. Similarly, scores on "Domineering-Submissive" constituted a peer-rating measure of Dominance, and scores on "Open-minded-Dogmatic" and "Flexible-Rigid" were summed for a measure of Open-mindedness.

The complete multitrait-multimethod matrix would consist of each of the three traits measured by each of the five methods. However, two trait-method units were not available: a rating scale measure of dominance and a checklist measure of dogmatism. Hence only 13 of the desired 15 measures were included.

Results

Since the relationships among the measures for the two fraternities were highly similar, the data for the two groups were combined. The matrix of intercorrelations for the total sample is presented in Table 1. The italicized diagonal values are validity coefficients. The solid triangles contain heterotrait-monomethod correlations, while the dotted lines indicate heterotrait-heteromethod triangles. A correlation of .23 or larger is significant beyond the .05 level. It should be noted that the checklist measure of self-esteem—distance between self and ideal-self concepts on the ICL—is interpreted such that the larger the distance, the lower the self-esteem. To make it consistent with other self-esteem measures, the signs of all correlations involving this distance index have been changed. Likewise, signs of correlations involving the Dogmatism scale have been reversed to make it consistent with the other measures of this trait, which are scored in the "Open-minded" direction.

Convergent Validity

Evidence for convergent validity may be determined by examining the italicized diagonal values. These are correlations between different methods of measuring the same trait. With respect

to the self-esteem measures, the CPI and Janis-Field scales and the self-ratings are highly intercorrelated (.67, .58, .60). Moreover, each of these measures is significantly related to peer ratings of self-esteem, although none of these relationships are very strong. Assessment of self-esteem by a discrepancy between self and ideal-self checklist responses appears to be something quite different. The self-ideal distance measure failed to correlate highly with any of the other indices of self-esteem, although its relationship with the CPI measure is significant.

The evidence for the convergent validity of dominance measures is more compelling in that none of the six validity coefficients is below .39. The correlation of .78 between the CPI and ICL dominance measures is particularly encouraging, especially since each is substantially correlated with peer ratings of this trait. Again, the self-ratings show adequate evidence of convergent validity.

Evidence for the convergent validity of measures of the dogmatism dimension is limited. The CPI Flexibility and (reflected) Dogmatism scales correlate .42, but neither measure is substantially related to the other indices of open-vs. closed-mindedness. Correlations of the Flexibility scale with self and peer ratings (.34, .25) are significant but not large, while correlations of Dogmatism with these ratings are meager (.11, .13). Self-ratings and peer ratings showed a significant but small relationship.

Discriminant Validity

Discriminant validity may be evaluated by examining three aspects of the correlation matrix: (1) whether a test's validity coefficients are higher than its correlation with other variables in heterotrait-heteromethod triangles, (2) whether a test's validity coefficients are higher than its correlations with other traits measured by the *same* method (i.e., its correlations in the heterotrait-monomethod triangle), and (3) the extent to which the same *pattern* of correlations among traits occurs in all of the heterotrait triangles. The evidence pertaining to the first two of these criteria is not impressive. That is, several of the off-diagonal values are of the same general magnitude as their corresponding diagonal elements (convergent validity coefficients). On this basis, then, one might conclude that evidence for discriminant validity is lacking.

Examination of the matrix in terms of the third criterion indi-

TABLE 1
Correlations among Measures of Self-Esteem, Dominance, and Dogmatism

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
<i>CPI</i>															
Self-Esteem	1														
Dominance	.67	2													
Flexibility	.10	-.04	3												
<i>Likert-type Scales</i>															
Janis-Field	.67	.52	.13	4											
—	()	()	()	.39	5										
Dogmatism	.09	.12	.42	()	()	6									
<i>ICL</i>															
Self-Ideal Distance	.26	.36	-.05	.20	()	.03	7								
Dominance	.65	.78	.06	.62	()	.16	()	8							
—	()	()	()	()	()	()	()	()	9						
<i>Self Ratings</i>															
Self Esteem	.58	.55	.13	.60	()	.17	.12	.60	()	10					
Dominance	.40	.54	.05	.36	()	.02	.09	.52	()	()	11				
Openmindedness	.04	-.02	.34	.14	()	.1	.15	.17	()	.06	()	12			
<i>Peer Ratings</i>															
Self-Esteem	.24	.42	.09	.23	()	.23	.02	.44	()	.28	.04	()	13		
Dominance	.25	.29	.05	.24	()	.03	.03	.41	()	.28	.09	()	()	14	
Openmindedness	.01	.10	.25	.09	()	.13	.01	.08	()	.03	.14	()	.86	.16	.39

Note.—The Dogmatism Scale and Self-Ideal Distance measures have been inverted to become measures of Openmindedness and Self-Esteem, respectively.
 $r = .23$ is significant at the .05 level.

cates, however, that this conclusion must be partially modified. This criterion is concerned with the consistency of interrelationships among traits in the heterotrait triangles. If the same traits are being assessed by the different methods, then the same pattern of correlations among the attributes should occur, regardless of the methods employed. In the ideal case in which only valid trait variance were assessed, the intertrait correlations would be the same in all heterotrait triangles and would indicate the degree of relationship among the traits themselves across persons. The correlations in Table 1 indicate a striking consistency of relationship among the attributes assessed in this study: self-esteem and dominance have substantial common variance, and each of these traits is unrelated to the openminded—closedminded dimension. The consistency of these relationships is evidenced in two ways. First, of the 20 instances of a correlation between a measure of self-esteem and a measure of dominance, 18 are significant, 11 of the coefficients being at least .40. The only nonsignificant relationships involved the ICL self-ideal distance, which apparently does not measure self-esteem as assessed by the other indices of this attribute. Second, of the 36 times a measure of open vs. closedmindedness is correlated with a measure of either self-esteem or dominance, only two are significant. Moreover, both of these cases (correlations between the Janis-Field and Dogmatism scales, and between peer-ratings of dominance and openmindedness) occur in monomethod triangles, where method variance is likely to artifactually inflate intertrait correlations.

Discussion

The results of this study indicate that the criteria for discriminant validity set forth by Campbell and Fiske (1959) are not entirely appropriate for the case of correlated attributes. In terms of two of these criteria, evidence for the discriminant validity of the self-esteem and dominance measures was generally lacking. Yet the consistency of intertrait correlations in the heterotrait triangles is impressive; indeed, this third criterion has rarely been met so thoroughly in other studies of this type in the literature. It is important to note that it is not simply the consistently high correlations between self-esteem and dominance that makes these data compelling; rather, it is this finding in conjunction with the consistent *lack* of relationship of either of these attributes with the measures of open-

mindfulness or flexibility. Clearly, attempts to differentially measure self-esteem and dominance have not been successful. This, however, does not require that we regard the two as the same construct; height and weight are highly correlated, yet are clearly distinct constructs. The usefulness of maintaining a distinction between two correlated constructs is based on their differential relationships with other variables (Kroger, 1968). Such a distinction has generally been made between self-esteem and dominance at the conceptual level. Empirical demonstration of the differential relationship of the measures of these constructs to other variables is needed.

Table 1 indicates that the four methods of assessing dominance were highly intercorrelated. The Dominance scale of the CPI appears to be one of the better scales in that inventory. It has consistently correlated well with peer ratings of dominance (Dicken, 1963; Gough, 1957) and has shown greater discriminant validity than some other CPI scales (Dicken, 1963). In the present study it was strongly correlated (.78) with the Dominance dimension of the ICL and showed substantial relationships with self and peer ratings of this attribute. The high correlations of ICL Dominance scores with the other measures (.78, .52, .41) support recent statements that the ICL may be a useful instrument and that further investigation into its psychometric properties is warranted (Bentler, 1965; Wiggins, 1968). Simple self ratings of dominance showed consistent and strong relationships with the other three methods (.54, .52, .50).

The five alternative methods of measuring self-esteem did not yield consistently high intercorrelations. However, the relationships among the CPI, Janis-Field, and self rating measures (.67, .58, .60) indicate that these three clearly form a cluster and are tapping the same attribute. Peer ratings of self-esteem were significantly correlated with each of these measures (.24, .23, .33) but the amount of variance they share with the highly-interrelated self-descriptive methods indicates that these ratings clearly are not a part of this cluster. The self-ideal distance measure was unrelated to peer ratings (-.02) and was only modestly correlated with the other three indices of self-esteem (.26, .20, .12). What we have, then, is a cluster of three highly intercorrelated self-descriptive measures, this cluster having partial overlap with both the peer ratings and the self-ideal

distance index; but these latter measures covarying with independent portions of the common variance shared by the three self-report measures.

The data of the present study do not permit a conclusive statement as to why the peer ratings did not correlate higher with the cluster of self-descriptive measures. One possibility is that the interrelationships among the three self-report methods were increased by common variance due to social desirability and other influences that enter into paper-and-pencil techniques. This interpretation would view peer ratings as representing a greater portion of true variance because of less susceptibility to artifactual influences. An alternative possibility has to do with the frame of reference from which the judgments are made. A person describing himself has access to a great deal of private experience (feelings, thoughts, etc.) not available to one who is judging another person. This interpretation might consider the self-report methods as possessing greater true variance than the peer ratings, which might then be considered as based primarily on a person's social stimulus value.

The Flexibility scale of the CPI correlated significantly with each of the other three measures of this attribute (.42, .34, .25); the only other significant convergent validity coefficient was between self and peer ratings of openmindedness. The Dogmatism scale was significantly correlated with the Flexibility scale, to an extent similar to that found in previous studies (Korn and Giddan, 1964; Rokeach, McGovney, and Denny, 1960), but was unrelated to either self or peer ratings. The peer rating measure was a combination of peer nominations on two dimensions—"Openminded-Dogmatic" and "Flexible-Rigid." Although these dimensions are clearly related, Rokeach et al. (1960) have argued for a conceptual distinction between rigid and closed-minded thinking. Indeed, the correlation between these two dimensions of peer nominations was .43, almost exactly the same as that between the (inverted) Dogmatism and Flexibility scales (.42). It might be argued, then, that one should not expect a high correlation between Dogmatism and the *combined* peer rating variable. However, the correlation between Dogmatism scores and ratings on the single dimension of "Openminded-Dogmatic" was only -.06. It should also be noted in this regard that self ratings made on a "Closed-vs. Open-minded" dimension were significantly correlated with the Flexibility scale (.34) but not with

the Dogmatism scale (.11). These findings question the ability of the Dogmatism scale to assess this attribute independently of a flexibility-rigidity construct.

It is interesting to note that none of the methods clearly outperformed the measures obtained by simple self-ratings on the attributes of interest. Not only were the convergent validity coefficients for the self ratings comparable to those for the other methods, but also the intercorrelations among the traits as assessed by this method (heterotrait triangles involving the self ratings) reflected the same interrelationships which so consistently appeared in the other heterotrait triangles. These findings lend further credence to Peterson's (1965) suggestion that it may be possible to adequately obtain personalistic information without employing the lengthy inventories traditionally used in such research. However, the conditions under which it is and is not appropriate to use this direct approach remain to be determined. One problem that seemingly would be important in such self ratings is social desirability. The extent to which these ratings are confounded with social desirability needs to be examined. Furthermore, the conditions and purposes of assessment may interact with social desirability in influencing these ratings. For example, when the assessment is for research purposes and subjects are assured of anonymity, these self ratings may be quite useable. On the other hand, when the assessment has important consequences for the individual (as in personnel selection), social desirability may become the dominant influence, thereby decreasing the validity and usefulness of these measures. It is necessary that these questions be examined before the encouraging results of both Peterson (1965) and the present study can be fully evaluated.

REFERENCES

- Bentler, P. B. Interpersonal Check List. In O. K. Buros (Ed.), *The Sixth Mental Measurement Yearbook*. Highland Park, N. J.: Gryphon Press, 1965.
- Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- Dicken, C. F. Convergent and discriminant validity of the California Psychological Inventory. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1963, 23, 449-459.
- Gough, H. C. *Manual for the California Psychological Inventory*. Palo Alto, California: Consulting Psychologists Press, 1957.

- Janis, I. L. and Field, P. B. Sex differences and personality factors related to persuasibility. In I. L. Janis and C. I. Hovland (Eds.), *Personality and persuasibility*. New Haven: Yale University Press, 1959.
- Korn, H. A. and Giddan, N. S. Scoring methods and construct validity of the Dogmatism Scale. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1964, 24, 867-874.
- Kroger, R. O. Conceptual and empirical independence in test validation: A note on Campbell and Fiske's "discriminant validity." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 383-387.
- LaForge, R. Research use of the ICL. Oregon Research Institute Technical Report, Eugene, Oregon, 1963.
- LaForge, R., Leary, T. F., Naboisek, H., Coffey, H. S., and Freedman, M. B. The interpersonal dimension of personality: II. An objective study of regression. *Journal of Personality*, 1954, 23, 129-153.
- Leary, T. F. *Interpersonal diagnosis of personality*. New York: Ronald, 1957.
- Loevinger, J. Objective tests as instruments of psychological theory. *Psychological Reports*, 1957, 3, 635-694.
- Meehl, P. E. The dynamics of "structured" personality tests. *Journal of Clinical Psychology*, 1945, 1, 296-303.
- Norman, W. T. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, 1963, 66, 574-583.
- Peterson, D. R. Scope and generality of verbally defined personality factors. *Psychological Review*, 1965, 72, 48-59.
- Rokeach, M. *The open and closed mind*. New York: Basic Books, 1960.
- Rokeach, M., McGovney, W. C., and Denny, M. R. Dogmatic thinking versus rigid thinking: An experimental distinction. In Rokeach, M. *The open and closed mind*. New York: Basic Books, 1960.
- Smith, G. M. Usefulness of peer ratings of personality in educational research. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 967-984.
- Wetzel, L. C. The convergent and discriminant validity of verbal personality measures. Unpublished Master's Thesis, University of Illinois, 1963.
- Wiggins, J. S. Personality structure. *Annual Review of Psychology*, 1968, 19, 293-350.

THE STABILITY OF INDIVIDUAL DIFFERENCES IN STRENGTH AND SENSITIVITY OF THE NERVOUS SYSTEM¹

FRANK H. FARLEY AND HERBERT H. SEVERSON

Wisconsin Research and Development Center for Cognitive Learning
University of Wisconsin

RESEARCH on individual differences (IDs) by Russian psychologists has differed in some major respects from the approach undertaken by most Western investigators. The latter have typically been concerned with such ID variables as intelligence, ability, and personality. These variables have usually been measured through more or less standard psychometric procedures such as paper-and-pencil tests, self-report inventories and projective techniques. Occasionally objective behavioral or apparatus measures are taken, or physiological indices recorded. The paper-and-pencil paradigm predominates, however, with scores on these tests usually being correlated with scores on other tests, or with measures of learning, perception, and so on. Where human learning is concerned, to consider one specific area of research, distinctions have been made between extrinsic and intrinsic IDs (Jensen, 1967). The former are considered to be sources of ID variance external to the learning process, e.g., anxiety, whereas the latter are seen as sources of ID variance internal or intrinsic to the learning process, e.g., susceptibility to interference in proactive and retroactive interference paradigms. In both cases, IDs are considered to contribute to learning.

A major Russian approach to IDs is represented by the work of

¹ This work was supported in part by the United States Office of Education, Department of Health, Education and Welfare under Contract OE5-10-154, Center No. C-03, and by the Graduate School Research Committee, University of Wisconsin.

Teplov (Gray, 1964), who has attempted to study the three Pavlovian dimensions of "strength," "equilibrium" and "mobility" of cortical excitation—inhibition in human Ss through the measurement of sensory and intersensory phenomena such as absolute visual thresholds (AVT) (held to reflect "sensitivity" of nervous functioning, and considered to be the inverse of the "strength" dimension), the effect on AVT of repeated peripheral visual stimulation, the effect on AVT of auditory stimulation, the effect on absolute auditory threshold (AAT) of visual stimulation, and so on. Through the use of extensive batteries of such measures, with subsequent factor analysis of these measures (Rozhdestvenskaya, Nebylitsyn, Borisova and Yermolayeva-Tomino, 1960), it has been possible to establish relatively unambiguous ID dimensions, particularly where the "strength" dimension is concerned. Pavlov (1927) considered "strength of nervous activity" or "strength of the nervous system" to be a fundamental ID dimension in man and other animals, referring to higher nervous activity and specifically to "strength of the excitatory process" and "strength of the inhibitory process." Teplov, who has undertaken the most extensive application of these notions to human behavior, has only considered, where strength is concerned, the "strength of the excitatory process." Although the Pavlovian hypotheses of neural activity held to underlie the ID dimensions would find little subscription in the West, the behavioral operations used in establishing the dimensions are acceptable procedures that can be judged on their own basis. The Russians have thus far been satisfied with reporting the dimensionality of sensory phenomena, and relating this often through experimental manipulations to putative processes of central excitation and inhibition. They have not attempted to delineate the representation of these IDs in learning, perception and motivation. The present senior author has been engaged in the development of a multi-dimensional approach to IDs in human learning which includes among its inputs a basic sensory dimensional analysis (SDA) alike in certain respects to that of Teplov, with, however, some antecedents in the (neglected) work of Galton. The next step beyond the factor analytic identification of sensory ID dimensions is being taken, which is the study of the contributions of these dimensions to learning phenomena. However, in attempting to integrate the Teplovian dimensions into our SDA work in human learning, it be-

came apparent that the stability of these dimensions required estimation. The "strength" dimension has been well established by the Russians (Gray, 1964) and accordingly, it was first considered as a potential contributor to an ID analysis of human learning. Teplov (1964) has reported that "sensitivity" of the nervous system, considered to be the inverse of strength, is well represented by AVT, which in the factor analysis by Rozhdestvenskaya et al. (1960) was found to have the highest loading of all measures on a strength dimension. Another measure of strength, the so-called Induction Method, had loadings on the strength dimension of from 0.52 to 0.74. This method involves, in all of its variants included in the Rozhdestvenskaya et al. (1960) study, the influence on the threshold taken to a point of light in peripheral vision of the presentation of an additional strong or weak light in the visual field. The effect here is that the threshold is *raised* by addition of a weak light and *lowered* by the addition of a strong light. One of the variants of the Induction Method has been called the "shape of the curve" measure (Rozhdestvenskaya, 1955), which relates, for each *S*, the sensitivity to the principal or main stimulus to intensity of the additional stimulus, with the latter usually being varied over a wide range. An abbreviated variation to this method introduced by Rozhdestvenskaya (1959) employed but one additional stimulus ($100 \times$ threshold) and determined the effect of this additional stimulus on sensitivity to the main stimulus. This is here called the "modified shape of the curve" index.

The present study was designed to obtain estimates of the stability of representative strength and sensitivity measures over a period of one month. The measures chosen were Rozhdestvenskaya's modified shape of the curve index and the AVT.

Method

Subjects

Fifteen university graduate students served as *Ss* (mean age = 26 years, range 22-30).

Apparatus

Thresholds were obtained with an NDRC model III adaptom-

eter.² The shape of the curve index involved the use of an additional light source peripheral to the main stimulus of the adaptometer. The adaptometer included a red fixation point (cross), directly below which at an angular distance of $2^{\circ} 17'$ was the main stimulus for visual threshold (dia. = $\frac{1}{2}$ in.), below which at an angular distance of $45'$ was the additional (peripheral) light source (dia. = $\frac{3}{2}$ in.). The angular height of all three stimuli was $3'$. These parameters were based on the report by Rozhdestvenskaya (1955). The adaptometer was powered by a six volt D. C. current with the stimulus light being controlled by a variable wedge neutral density filter. Color was controlled with a Kodak 540 mu. filter.

Procedure

Ss were dark adapted while wearing red lucite goggles in a semi-dark (15 watts illumination) 6 ft. \times 6 ft. sound-reduced room for 30 minutes. Following this, each S was moved to a similar testing room (6 ft. \times 6 ft.) where, while seated in the threshold testing apparatus in total darkness, he was dark adapted for a further 10 mins. For actual testing, the S's chin rested in a Bausch and Lomb Model BA5372 chin rest, such that his eyes were 24 in. from the adaptometer face directly in line with the center of the main light source. Binocular viewing was used. The test patch was $\frac{1}{2}$ in. in diameter and located 20° angular distance below the fixation point, with the angular size of the aperture being $1^{\circ} 30''$. The S was located in a totally darkened cubicle within the experimental room itself, while the E was located outside this cubicle. The interior of the cubicle, including apparatus, was entirely matt black in finish.

The S was instructed to fixate upon the red fixation cross. He was told that the E would deliver an auditory cue (pure tone) following which the main visual stimulus directly below the fixation point would be presented for 1 sec. He was to respond with a simple yes or no as to whether or not the light was perceived. A modified method of limits was employed, with a criterion of two consecutive yes-responses on the descending series and two consecutive no-responses on the ascending series. The mean of the two series was taken as the AVT.

² The authors would like to thank Prof. F. A. Mote of the University of Wisconsin for the long-term loan of this apparatus.

The second phase of the study obtained the modified shape of the curve index. This consisted of introducing the additional, peripheral, visual stimulus at the light source located directly below the main test patch, at a light value $110 \times$ the threshold for all Ss. The S's threshold to the main stimulus was then taken once in the presence of the additional light source. The identical procedure to AVT was used to obtain the threshold. The S fixated on the red fixation point throughout.

Both the AVT and modified shape of the curve data were expressed in log luminance units derived from the neutral density wedge at the main light source. Periodic checks with a photomultiplier indicated the luminance to be constant at given wedge locations.

The experimental procedure took approximately 15 minutes following dark adaptation. The Ss were tested between 9 a.m. and 2 p.m., with no Ss being tested during lunch hour. The Ss were asked not to drink coffee or other high caffeine drinks on the day of testing, as caffeine has been shown to influence the measures employed in the study (Gray, 1964) and has been used extensively in the Russian work on strength as a variable held to influence 'cortical excitability.' All Ss tested had normal visual acuity with no history of eye problems.

The retest session (Session 2) for the stability estimate was undertaken exactly one month after the first test (Session 1), at the same time of day, with the same E and under identical conditions. The same procedures were used in Session 2 as were used in Session 1, including the dark adaptation period.

Results and Discussion

Where the AVT is concerned, the mean threshold for Session One was 5.28 log luminance units, whereas the mean on Session Two was 4.10 log luminance units. A *t* test of this difference indicated it to be highly significant ($p < .001$), demonstrating that the mean AVT had significantly decreased from the first to second testing. A scatterplot of the two sets of scores indicated no curvilinearity of relationship, but rather a strong positive linear relationship. The product-moment correlation coefficient was .91 ($p < .001$) indicating that approximately 83 per cent of the variance in AVT on Session Two was accountable in terms of Session One performance.

Turning to the modified shape of the curve index, the mean threshold for Session One was -1.16 log luminance units, whereas the mean on Session Two was $-.38$ log luminance units. This difference was significant by t test ($p < .02$) indicating a significant decrease in the mean modified shape of the curve value from first to second testing. A scatterplot of the two sets of scores indicated no curvilinearity of relationship, but rather a positive linear relationship. The product-moment correlation was $.61$ ($p < .02$) demonstrating that approximately 37 per cent of the variance in the modified shape of the curve measure on Session Two was accountable in terms of Session One performance.

The major finding of the study is the marked stability of IDs in AVT, and the moderate stability of IDs in the modified shape of the curve index. The AVT stability estimate was as high or higher than is usually found with intellectual or personality measures, and clearly meets the primary requirement for use in ID studies. On the basis of the Rozhdestvenskaya, et al. (1960) factor analysis where the highest loading on a strength factor was AVT, it might be suggested that taking into account the present results, this measure should be strongly recommended for use as an index of strength. It possesses high factorial validity and exceptionally high reliability. In Teplovian theory it is a measure of sensitivity and is thus the inverse of strength. This, of course, does not change the fact that with its high loading on the strength dimension it can be used in practice as a measure of strength. The high loading of AVT on the strength factor may, of course, be in part due to its very high reliability. The lower loadings of many of the other putative measures of strength in the Rozhdestvenskaya, et al. (1960) factor analysis may have been due to lower reliabilities, as found with the present modified shape of the curve index. The latter measure as well as many remaining strength measures are considerably more complicated procedurally than the simple AVT. The introduction of additional sources of stimulation into the threshold testing situation, as well as the longer period of time required to complete the task, with accompanying postural fatigue or boredom, might be expected to lead to greater intra-individual response variability with a possible consequence in attenuated reliability estimates. The modified shape of the curve reliability estimate of $.61$, however, is not discouraging enough to forsake this measure. Indeed, the safest

approach to the measurement of as nebulous a concept as strength is through multiple-indices, among which the AVT and modified shape of the curve should be numbered.

Summary

The stability over one month of representative measures of strength ("modified shape of the curve") and sensitivity (absolute visual threshold—AVT) of the nervous system as determined from the research and theory of Teplov and associates at the University of Moscow was estimated on 15 Ss. The stability estimate for the "modified shape of the curve" measure was .61 ($p < .02$) while that for the AVT was .91 ($p < .001$). The results were discussed in regard to the Russian factor analytic identification of a dimension of strength, and in relation to the choice of strength measures for further research. A multiple-indice approach was recommended.

REFERENCES

- Gray, J. A. *Pavlov's typology*. New York: Macmillan, 1964.
- Jensen, A. R. Varieties of individual differences in learning. In R. M. Gagné (Ed.) *Learning and individual differences*. Columbus, Ohio: Merrill, 1967.
- Pavlov, I. P. *Conditioned reflexes*. New York: Oxford University Press, 1927.
- Rozhdestvenskaya, V. I. An attempt to determine the strength of the process of excitation through features of its irradiation and concentration in the visual analyzer. *Voprosy Psikhologii*, 1955, 3, 90-98.
- Rozhdestvenskaya, V. I. Strength of nerve-cells as shown in the nature of the effect of an additional stimulus on visual sensitivity. In B. M. Teplov (Ed.) *Typological features of higher nervous activity in man*, Vol. 2. Moscow: Akad. Pedagog. Nauk RSFSR, 1959.
- Rozhdestvenskaya, V. I., Nebylitsyn, V. D., Borisova, M. N., and Yermolayeva-Tomina, L. B. A comparative study of various indices of strength of the nervous system in man. *Voprosy Psikhologii*, 1960, 5, 41-56.
- Teplov, B. M. Problems in the study of general types of higher nervous activity in man and animals. In J. A. Gray (Ed.) *Pavlov's typology*. New York: Macmillan, 1964.

A REVISED PROCEDURE FOR THE ANALYSIS OF BIOGRAPHICAL INFORMATION

WILLIAM H. CLARK AND BRUCE L. MARGOLIS¹

Case Western Reserve University

WHEN examining personal history or biographical information blanks (BIBs) as potential predictors in different research projects, the authors noted a slight discrepancy in the item analysis procedures described by Stead and Shartle (1940) and England (1961).

Briefly, these procedures include calculation of response frequencies and differences in proportions between criterion groups and derivation of "net weights" for response categories from the differences in proportions according to Strong's tables. The net weights derived from Strong's tables range from -28 to $+28$. Since this range of net weights may appear to imply greater predictive ability than is in fact true, net weights are converted to "assigned weights" reflecting a simple positive, negative, or absence of relationship with the criterion. Only those items achieving a specified net weight, either positive or negative, are given assigned weights and included in the final scoring system for the BIB.

It is at this point that the discrepancy between the two procedures appears. To differentiate between the criterion groups, and thus to be included in the final BIB scoring system, any response must have a net weight whose absolute value is two or greater, according to Stead and Shartle (1940, p. 255). England, on the other hand, requires an absolute value of four or greater for acceptance of a response as differentiating between the groups. England's only comment regarding this more stringent requirement is given a brief

¹ Now at A. T. Kearney & Company, Inc. Chicago and U.S. Public Health Service, Cincinnati, respectively. The authors wish to express their appreciation to Dr. Douglas Schultz for his comments on early drafts of this paper.

footnote: the requirement "is modified from that suggested by Strong as to weight fewer chance differences between the weighting groups" (England, 1961, p. 25). Not only is this difference in requirements unexplained, but also, neither procedure explains the statistical basis for consideration of any net weight as discriminating.

Using these methods in an item analysis of BIB's developed for two independent selection research projects, the authors also found themselves unable to determine the probability of Type I error. To determine the reasons for the discrepancy between Stead and Shartle's and England's approach and also to search for any established method of determining Type I error probabilities, the authors reviewed Strong's original development of the item analysis procedure (Strong, 1926).

Strong utilized Kelly's formula (Cowdery, 1926) to obtain b values determined from phi coefficients. In using b values Strong was setting up a multiple regression equation between the item predictors and the criterion. His rationale, however, for the jump from net weights to assigned weights is unclear. The assigned weights, it would seem, should be based upon the probability that the response is discriminating. The probability that a given phi coefficient is not a chance deviation from zero can be determined but a more direct approach is available to determine whether or not an item is discriminating. For example, a simple t test can permit the determination of the probability associated with differential response frequencies for different criterion groups (such as males and females or high and low performers.)²

Positive or negative weights can be given to responses which are shown to discriminate significantly and zero weights can be given

² The authors used the following formula:

$$t = \frac{p^1 - p^2}{\sqrt{\frac{p^1 q^1}{n^1} + \frac{p^2 q^2}{n^2}}}$$

where:

p^1 = proportion of one criterion group choosing the given response.

p^2 = proportion of other criterion group choosing the given response.

$q^1 = 1 - p^1$, $q^2 = 1 - p^2$.

n^1 = size of first group responding to entire question.

n^2 = size of the second group.

(Burt, 1942, p. 329).

to responses which do not discriminate between the groups. Correlation methods may also be used to obtain similar results.

Strong appears to have developed his method to permit researchers to circumvent the lengthy, tedious calculation of probabilities associated with given phi coefficients or t values. Using Strong's method and tables, item weights are obtained from examination of proportions of item responses given by the criterion groups. With the availability of high speed data processing, however, it becomes a very simple process to determine the precise probability of Type I errors. The authors felt the determination of t values and the probability associated with them for each response would permit a more clear and meaningful analysis of BIB responses.

There are other advantages to the calculation of t values for each item. Strong and England have both insisted on large sample sizes (greater than 100) in both criterion groups and their tables assume this requirement by taking into account the sample size and any variations in size among items. Calculation of the exact probabilities allows the researcher to use his own criteria and judgment for inclusion or exclusion of items in the final weighting system.

Empirical tests of the comparability of Strong's method and the calculation of t values showed a high degree of similarity between the two approaches. Using one set of data, the net weights derived for Strong's method were correlated with t values calculated from the same data. A Spearman rank-order correlation coefficient of .91 was obtained. Using a different set of data, a Pearson product-moment correlation coefficient of .85 was obtained between net weights found in Strong's tables and actual probabilities of Type I errors determined from t values.

With the advent of high-speed processing methods, the somewhat obscure and poorly explained correlational shortcuts originally developed by Strong may be replaced by a direct, hypothesis-testing approach which takes account of varying sample sizes and provides exact probabilities, thus allowing the researcher greater freedom to exercise his judgment.

REFERENCES

- Burt, H. E. *Principles of employment psychology*, (Revised ed.). New York: Harper, 1942.
- Cowdery, K. M. Measurement of professional attitudes: Differ-

- ences between lawyers, physicians and engineers. *Journal of Personnel Research*, 1926, 5, 131-141.
- England, G. W. *Development and use of weighted application blanks*. Dubuque, Iowa: Wm. C. Brown Company, 1961.
- Stead, W. H. and Shartle, C. L. *Occupational counseling techniques: Their development and application*. New York: American Book Company, 1940.
- Strong, E. K., Jr. An interest test for personnel managers. *Journal of Personnel Research*, 1926, 5, 194-203.

VALIDITY STUDIES OF ACADEMIC ACHIEVEMENT

WILLIAM B. MICHAEL, Editor
University of Southern California

JOAN J. MICHAEL, Assistant Editor
California State College, Long Beach

<i>An Empirical Validity Study of the Assumptions Underlying the Structure of Cognitive Processes Using Guttman-Lingoes Smallest Space Analysis.</i> H W. STOKER AND R. P. KROFF	469
<i>Validity of Taxonomic Tests.</i> I. LEON SMITH	475
<i>Measurement of College Achievement by the College-Level Examination Program.</i> AMIEL T. SHARON	477
<i>Concurrent Validity of a Literature Test in Relation to Selection of Persons for Graduate Study in English.</i> JOSEPH P. SCHNITZEN AND JOHN A. COX	485
<i>Predicting Quality Point Averages in Master's Degree Programs in Education.</i> JERRY B. AYERS	491
<i>Another Contribution to Estimating Success in Graduate School: A Search for Sex Differences and Comparison between Three Degree Types.</i> DAVID A. PAYNE, ROBERT A. WELLS, AND ROBERT R. CLARKE	497
<i>Prediction of Choice of and Success in Agriculture as a College Major.</i> JAMES M. RICHARDS, JR.	505
<i>Use of the ROTC Qualifying Examination for Selection of Students to Enroll in Advanced Courses in ROTC as Juniors.</i> THOMAS M. GOOLSBY, JR. AND DONALD A. WILLIAMSON	513
<i>A Note on the Predictive Validity of the Cooperative Algebra III.</i> CLIFFORD B. TATHAM AND ELAINE J. TATHAM	517
<i>Statistical Analysis of Three Critical Thinking Tests.</i> JOHN FOLLMAN, WILLIAM MILLER, AND ELDON BURG	519

- Cross-Validation of the Orleans-Hanna Algebra Prognosis Test and the Orleans-Hanna Geometry Prognosis Test.* JOANNE M. LENKE, HAROLD F. BLIGH AND BERNARD H. KANE 521
- A Comparison of the D-48 Test and the Otis Quick Score for High School Dropouts.* BRAD S. CHISSOM AND RALPH LIGHTSEY 525
- The Relative Predictive and Construct Validities of the Otis-Lennon Mental Ability Test, the Lorge-Thorndike Intelligence Test, and the Metropolitan Readiness Test in Grades Two and Four: A Series of Multivariate Analyses.* BARTON B. PROGER, JOHN R. MCGOWAN, ROBERT J. BAYUK, JR., LESTER MANN, RUTH L. TREVORROW, AND EDWARD MASSA .. 529
- The Relationship of Average Scores on Intelligence and Reading Tests to Percentages of Minority Group Students in Elementary Schools and High Schools in a Large Metropolitan Area.* WILLIAM B. MICHAEL, ROBERT A. SMITH, AND YOUNG B. LEE 539
- The Development of a Measure of Vocational Maturity.* BERT W. WESTBROOK, JOSEPH W. PARRY-HILL, JR. AND ROGER W. WOODBURY 541
- A Partial Redefinition of the Factorial Structure of the Study Attitudes and Methods Survey (SAMS) Test.* WILLIAM B. MICHAEL, YOUNG B. LEE, JOAN J. MICHAEL, ORA HOOKE, AND WAYNE S. ZIMMERMAN 545
- A Factor Analysis of the CPI and EPI.* ROBERT D. ABBOTT .. 549
- The Reduced Size Rod and Frame Test as a Measure of Psychological Differentiation.* TED NICKEL 555

PROVISION FOR PUBLICATION OF VALIDITY STUDIES OF ACADEMIC ACHIEVEMENT

Early in the life of this journal it became evident that the prediction of academic achievement is by far the most popular area of research in the measurement field. It also became apparent that unless heroic measures were taken, the journal might easily be practically monopolized by this subject. The heroic measure resorted to for a while was simply not to publish any studies on the prediction of academic achievement.

In the course of time, it became evident that the solution hit upon was too drastic. After all, it is important that validity reports be available, at least in condensed form, to educational and personnel psychologists and to school counselors who wish to evaluate the relative merits of the various instruments available for the prediction of academic achievement. Furthermore, it appears that a substantial amount of validity data cannot be conveniently communicated to professional workers in the field of measurement unless provision is made for publication in a professional journal.

In the light of this situation, the policy has been adopted of publishing a section devoted to such studies in the form of extra pages for which the authors bear most of the publication costs. This policy allows publication of the usual number of pages on other subjects in the measurement field. The charges consist of thirty-five dollars per page of running text plus any extra costs which may be involved in the composition of tables, figures, and formulas. Authors are furnished one hundred off-prints without extra charge.

Preference will be shown for manuscripts of fewer than 1200 words, with no more than six references and containing two or fewer tables each of no more than $8\frac{1}{2}'' \times 11''$ elite typed page—making six printed pages. Any manuscript exceeding 3000 words, 12 references, and four tables or figures equivalent to three $8\frac{1}{2}'' \times 11''$ typed pages will be automatically returned, as 12 printed pages will be the maximum total number of pages for any article to be published in this section.

The Validity Studies of Academic Achievement Section is published twice a year, once in the Summer issue and again in the Winter issue, for which the closing dates for receiving manuscripts are November 30th and May 30th, respectively.

Two copies of the manuscripts should be sent to:

Dr. William B. Michael
325 Callita Place
San Marino, California 91108.

AN EMPIRICAL VALIDITY STUDY OF THE ASSUMPTIONS UNDERLYING THE STRUCTURE OF COGNITIVE PROCESSES USING GUTTMAN-LINGOES SMALLEST SPACE ANALYSIS

H. W. STOKER AND R. P. KROPP
Florida State University

THE authors of the *Taxonomy of Educational Objectives: Cognitive Domain* (Bloom, 1956) regarded the structure of the cognitive processes to be cumulatively hierarchical. That is, the major classes—knowledge, comprehension, application, analysis, synthesis, and evaluation—could be placed on a continuum such that each successive class includes all behaviors represented by the class or classes preceding it. The authors also suggest that these classes of abilities, behaviors, or processes transcend content. Previous studies (Kropp and Stoker, 1966; Smith, 1968) investigated the construct validity of these assumptions of cumulative hierarchy and transcendence of process separately. In this study the Guttman-Lingoes smallest space analysis (SSA-1) technique is used in an attempt to examine and validate both assumptions simultaneously.

Guttman (1954) discussed the concepts of the simplex, circumplex, and the radex. If a perfect simplex occurs, certain partial correlations will vanish. If

$$r_{ik} = r_{ij} \cdot r_{jk} \quad \text{where } i < j < k$$

then $r_{ik,j} = 0$ for all $i < j < k$. When this relationship is satisfied, the resultant correlation matrix will have the largest values along the upper-left, lower-right diagonal followed by the next larger values in the adjacent diagonal and the smallest values in the upper-right and lower-left corners of the matrix. The relationship is such that the col-

umn sums of the correlation matrix for a perfect simplex will be smallest at the extremes and largest for the central columns. Assuming that the correlations arose from a set of tests, the existence of a simplex would imply an ordering of the tests along some dimension. The assumption of cumulative hierarchy in the taxonomy implies that a set of tests designed to measure the mental processes should exhibit the simplicial structure along a complexity dimension.

If a uniform, perfect, additive circumplex exists, then

$$t_{ij} = \begin{aligned} & c_{i,i} + c_{i+1,i} + \cdots + c_{i+m-2,i} & (j \geq n - m + 1) \\ & c_{1,i} + c_{2,i} + \cdots + c_{i-n+m-1,i} \\ & \quad + (c_{i,i} + \cdots + c_{n,i}) & (j > n - m + 1) \end{aligned}$$

where t_{ij} represents a test (t_i) with elementary additive components (c_i). Assuming that all elementary components are uncorrelated ($r_{c_p c_q} = 0$ for $p \neq q$) and considering the special case where the components have equal variance and where $m \geq n/2$, then:

$$r_{ijk} = \begin{aligned} & 1 - \frac{k-j}{m} & 0 \leq k-j < n-m \\ & 1 - \frac{n-k+j}{m} & n-m \leq k-j < n. \end{aligned}$$

The characteristics of a matrix of intercorrelations which satisfy these restrictions are that the column totals will be equal and each row of the table will have the same entries as the preceding row, but moved one column to the right, the end one moving to the beginning. A matrix exhibiting these characteristics was referred to as a "circulant" by Guttman (1954). If process transcends content, as suggested by the authors of the Taxonomy, then tests at the same process level should have higher correlations than tests of different process levels, implying that one should find a circumplex when dealing with taxonomy-type tests.

When the correlations do not meet the rigorous requirements for either a simplex or a circumplex but the matrix exhibits the characteristics of one or the other, the matrix may be classed as a quasi-simplex or a quasi-circumplex.

In a radex, both types of ordering occur simultaneously and two dimensions should suffice for representing the relationships. In this

study, the relationships between tests of differing process levels should decrease as the distance between them on the complexity dimension increases; tests of the same process level should have higher intercorrelations than tests of different process levels. These relationships will be reflected by the distance between the test when plotted in two-space. [For additional background, the reader is referred to two other contributions by Guttman (1964, 1968)].

Procedure. The data analyzed were from a study conducted by the authors, Kropp and Stoker (1966). Taxonomy-type tests, designed to measure the six levels of mental processes as defined in the Taxonomy, were constructed for each of four contents. Tests were based on reading passages presented to the students and available to them while responding to test items for all but the synthesis and evaluation levels. Reading passages were selected on the basis of ease of comprehension, interest value, and unfamiliarity of the material to the students. The tests, which were entitled, "Atomic Structure," "Glaciers," "Lisbon Earthquake," and "Stages of Economic Growth," are the same ones used in the Smith studies (1968).

Tests of knowledge, comprehension, application, and analysis each contained 20 multiple-choice questions. Each synthesis test consisted of five free-response questions, scored zero to four; each evaluation test consisted of 10 free-response items, scored zero to two. Thus, a maximum of 20 points could be earned on each test. Tests were administered to students in grades nine through 12. Correlation coefficients were derived for each grade level and for all grades combined. The number of scores contributing to each coefficient varied, but within grade, the number exceeds 750 in each case.

The five 24×24 intercorrelation matrices formed by intercorrelating scores within and over grades were subjected to an analysis using the Guttman-Lingoes SSA-1 computer program. The solution described below is the one obtained by hypothesizing that two dimensions would represent the data. Other solutions were tried, but the use of two dimensions appeared to yield the best solution.

If the assumptions of cumulative hierarchy and transcendence are tenable, the graphs produced by the well known Guttman-Lingoes SSA-1 program should provide evidence of a radex. One dimension on the graph should represent complexity of process and the other the generality of process over content. Hence, the combination of the specially constructed tests and the analysis technique provides a

means by which the construct validity of the underlying assumptions can be examined.

Results. Graphs representing the data for grades nine through 12 and all grades combined were constructed. Figure 1 contains the solution for all grades combined. The solution for grades nine through 12 yielded different graphs, but with some similarities. The graph for all grades combined is provided for illustration of the results of such a solution. The solution represented by the figure is the one resulting from the specification of two dimensions. Coefficients of alienation, which refer to the spread in the scattergram yielded by the program, were approximately .20. These coefficients indicate the degree to which

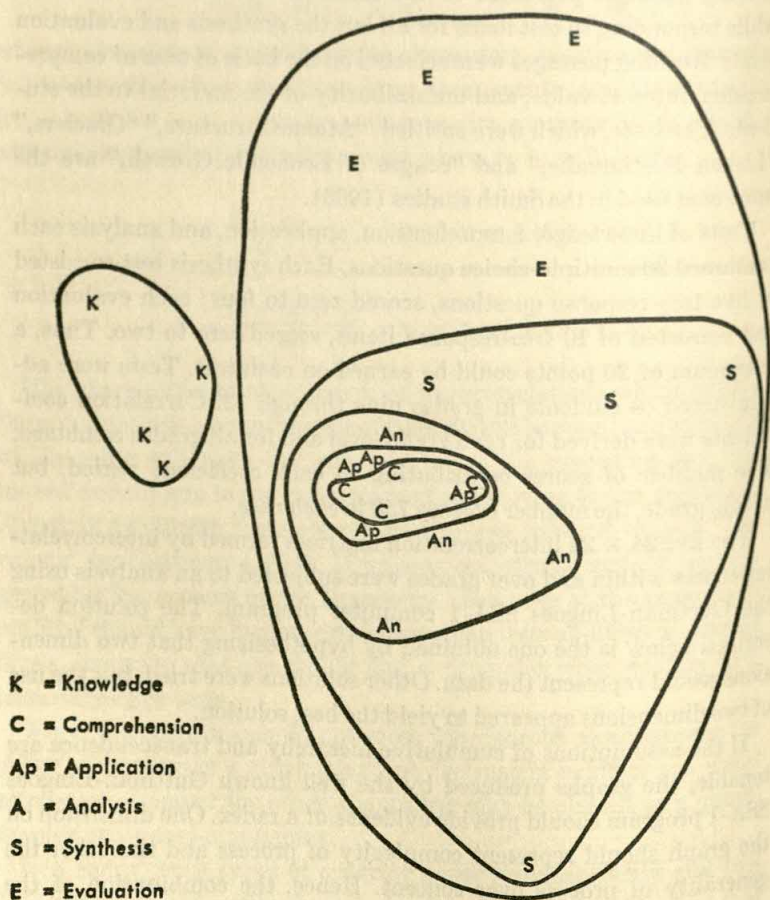


Figure 1. GL-SSA-1 solution for two dimensions all grades combined.

$r_{ij} = f(d_{ij})$, where f is a monotonically decreasing function. A zero value would represent a perfect fit.

In all five graphs there was some evidence of the existence of a radex. The "center" for grades 9, 10, 12 and all grades, was the set of tests measuring comprehension with application, analysis, synthesis, and evaluation radiating out in order from this center. For grade 11, the "center" was a mixture of comprehension and application.

The notable exception of the hypothesized outcome, in each figure, was the set of tests designed to measure knowledge. The grouping of knowledge tests for each grade reflects the high intercorrelation of the knowledge tests coupled with a somewhat random distribution of correlations of knowledge tests with other tests. The fact that these tests were designed to maximize the number of individuals receiving a perfect score probably accounts for the pattern of correlations obtained and the separation in two-space of these tests from the others.

A somewhat similar patterning appeared in the graphs for grades 11, 12 and for all grades, with respect to the evaluation tests. These tests proved to be extremely difficult, yielding markedly skewed distributions. While there was some correlation among the evaluation tests (average $r = .50$), there was no definite pattern in the correlation of the evaluation tests with tests at other levels. The lack of a pattern, coupled with the intercorrelation among the evaluation tests probably explains the grouping pictured in these graphs.

REFERENCES

- Bloom, B. S. (Ed.) *Taxonomy of educational objectives. Handbook 1: Cognitive domain*. New York: McKay, 1956.
- Guttman, L. A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the behavioral sciences*. Glencoe: The Free Press, 1954.
- Guttman, L. The structure of interrelations among intelligence tests. Proceedings of the 1964 Invitational Conference on Testing Problems. Princeton: Educational Testing Service, 1964.
- Guttman, L. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, 1968, 4, 469-505.
- Kropp, R. P. and Stoker, H. W. The construction and validation of tests of the cognitive processes as described in the *Taxonomy of educational objectives*. Cooperative Research Project No. 2117. U. S. Office of Education, 1966.
- Smith, R. An empirical examination of the assumptions underlying the taxonomy of educational objectives: Cognitive Domain. *Journal of Educational Measurement*, 1968, 5, 125-128.

VALIDITY OF TAXONOMIC TESTS

I. LEON SMITH

University of Cincinnati

In order to validate tests derived from Bloom's (1956) *Taxonomy of Educational Objectives*, it is necessary to base items on content for which students have relatively the same mastery so that score variability will reflect differential mastery of the cognitive processes. The latter are defined as Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation.

In a recent attempt (Kropp, Stoker, and Bashaw, 1966), the method for controlling content mastery involved the introduction of material, at the time of testing, which was assumed to be equally unfamiliar to all students. This was accomplished by presenting the students with a reading passage and basing the taxonomic items on its content. This study reports on the use of this type of response measure in an effort to hold content knowledge constant.

Method. The Kropp et al. (1966) test of the Stages of Economic Development (SED) and the Social Studies (SS) subtest of the Stanford Achievement Test: High School Battery were administered to 141 eleventh-grade students. The IQ range of the group was 67 to 143 as measured by the Lorge-Thorndike Intelligence Test, Level G, Form 1.

Results and discussion. Support for the use of unfamiliar material as a method for controlling content knowledge would be indicated if the relationships between the SED tests and the measure of subject-matter mastery (SS) were low and insignificant. However, in view of the substantial correlations between the SED tests and the SS measure as shown in Table 1, there is the strong suggestion that performance on each taxonomic test depends heavily on relevant past

TABLE 1
Reliabilities and Correlations among Tests

	SS	Reliability*
SED		
Knowledge	.73*	.84
Comprehension	.73*	.81
Application	.64*	.86
Analysis	.71*	.74
Synthesis	.72*	.72
Evaluation	.60*	.71
SS		.88

* $p < .05$.

* Reliability estimates obtained from test manual.

achievement. It appears that content knowledge is not controlled by the presentation of unfamiliar material. Thus, the interpretation of score variability in terms of differential mastery of the cognitive processes may not be warranted. Since the use of any response measure is likely to permit relevant past knowledge and experience to be transferred to the testing situation, the validity of the process levels of the *Taxonomy* may be difficult to establish. However, the results are encouraging in the sense that the measure of standardized achievement does seem to tap all of the behaviors postulated by the *Taxonomy* as measured by the SED test.

REFERENCES

- Bloom, B. S. (Ed.) *Taxonomy of educational objectives*. New York: Longmans, Green, 1956.
- Kropp, R. P., Stoker, H. W., and Bashaw, W. L. *The construction and validation of tests of the cognitive processes as described in the taxonomy of educational objectives*. Cooperative Research Project #2117, U. S. Office of Education, Institute of Human Learning, and Department of Educational Research and Testing, Florida State University, 1966.

MEASUREMENT OF COLLEGE ACHIEVEMENT BY THE COLLEGE-LEVEL EXAMINATION PROGRAM

AMIEL T. SHARON

Educational Testing Service

THE General Examinations (GEs) of the College-Level Examination Program (CLEP)¹ are intended to provide a comprehensive measure of undergraduate achievement in five basic areas of liberal arts: English, natural sciences, humanities, mathematics, and social sciences-history. The tests are not designed to measure advanced training in any specific discipline but rather to assess a student's knowledge and comprehension of basic facts, concepts, and principles in each of the five subjects. The content covered by the GEs is similar to the content included in the program of study required of many liberal arts students in the first two years of college. It has been developed by committees of specialists in each of the subject-matter fields. The committees work with test specialists in defining the topics to be covered, reviewing the test specifications, and suggesting and reviewing test questions.

In addition to being used for granting college credit or placement for military service experiences, television and correspondence courses, and independent study, the GEs are used for a variety of other purposes at collegiate institutions. They are employed for guiding students into appropriate curricula of study; admitting and placing transfer students; assessing student growth in various curricula; and selecting students for upper division studies. Many colleges and universities are also using the examinations for self-study and for research on specific questions about types of students, courses, or

¹ The CLEP, which is sponsored by the College Entrance Examination Board, includes both the General and Subject Examinations.

curricula. The questions which are asked range from "How do our sophomores compare with those at other colleges in terms of their liberal arts education?" to "Does exposure to our liberal arts courses result in greater knowledge as measured by these tests?"

The most common procedure for demonstrating the appropriateness or validity of achievement tests, such as the GEs, is by means of content validation. The test content is developed systematically to be representative of the subject matter to be measured. In addition, empirical procedures such as item analysis aid the test specialists in deciding on which items to include in the examinations. Since the GEs have been constructed by rigorous procedures of content validation described elsewhere (ETS, 1965), the present report focuses on the empirical validity of the tests.

Two different types of empirical validity will be discussed: criterion-related validity and construct validity. Criterion-related validity is useful for prediction of future performance and assessment of current achievement level. The criterion-related validity of the GEs will be described in terms of the relationship of the tests to college grades. Although the grade-point average (GPA) criterion has been criticized for being unstable and for failing to reflect certain desirable types of student traits such as ethicality, openmindedness, altruism, maturity, and self-insight, its ready availability has promoted its use as a criterion of college success by many researchers.

Unlike criterion-related validity, construct validity aims to increase understanding of the educational or psychological attributes measured by a test. It requires the gathering of information from a variety of sources. The construct validity of the GEs will be described by the inferred effect of college instruction on test performance and by the differential performance of various types of students on the examinations. The possibility of the examinations being inappropriate to certain types of students, a topic closely related to validity, will also be discussed.

Criterion-related validity. Positive correlations between the GEs and overall GPA, in most cases overall sophomore GPA, have been reported in studies conducted at six universities. Since GPA and the scores on GEs were collected simultaneously in these studies, these correlations represent the concurrent validity of the examinations. Invariably the English Composition Test was found to be the most valid one, with a median coefficient of .46. The rank order of the va-

validity coefficients of the four other examinations was not consistent across the different studies. Median validities were Natural Sciences .40, Humanities .40, Social Sciences-History .36, and Mathematics .30. These correlations indicate that there is a moderately positive, but far from perfect, relationship between the tests' scores and grades. This result is not too surprising, since grades in many courses are based on objective tests similar in content and format to the GEs. Nevertheless, these results suggest that the tests can be used legitimately for granting course credit or placement in college.

The correlations between the GEs and grades in subjects corresponding to each test are in general no higher than the test's correlations with overall GPA. This conclusion is based on studies conducted at two universities. A probable explanation of these results is that overall GPA is more reliable than subject GPA because it is based on a larger number of courses.

The validity of the GEs when taken at the end of the sophomore year, for predicting junior or junior/senior grades, is significantly lower than the concurrent validity of the tests. Median validity coefficients computed on the basis of three studies were English Composition .36, Humanities .28, Natural Sciences .27, Social Sciences-History .26, and Mathematics .15. Again, the English Composition and the Mathematics Tests appear to be the most and least valid tests respectively. The reason for the low validity of the Mathematics Test may be that mathematics plays a very minor role in courses taught in the last two years of college. The finding that the predictive validities of the GEs are lower than their concurrent validities indicates that the tests are less useful for guidance or prediction of success in upper-level studies than they are as measures of current achievement level.

Construct validity. Construct validity indicates the extent to which a test can be said to measure a trait or a theoretical construct. It also refers to the ability of a test to yield reasonable results, consistent with expectations. For example, a scholastic achievement test should yield higher scores for those who have more education than for those who have less education; history majors should score higher on a history test than biology majors; and students should have higher scores on an algebra test after taking an algebra course than before taking the course.

There are two reasonable expectations or implicit assumptions

underlying the College-Level Examination Program which have implications for the construct validity of the GEs:

1. There is a gain in knowledge resulting from college instruction which can be measured by an examination.
2. The examinations employed to measure gain in knowledge are appropriate to the courses taught at the colleges.

These assumptions have implications which extend beyond those underlying the coefficient of correlation. In demonstrating that there is a positive correlation between test scores and grades no claim can be made that test scores or grades are affected by instruction. In order to determine whether a change in test performance is influenced by college instruction, it is necessary to administer the test *before* and *after* the course of instruction. Also required would be the testing of one or more control groups (to which students would be randomly assigned) who would not receive instruction appropriate to the test or any instruction at all. Without a control group, any gain achieved on the examinations could be interpreted as resulting from intellectual growth rather than from a specific course of study. Unfortunately, it is difficult to have control groups in educational research. The notion of "manipulating" the learning of students for the sake of research is anathema to many educators. None of the studies which employed a "before-after" design to study score gain on the GEs employed a control group.

Harris and Booth (1969) reported on gains made on the GEs from the first to the sixth quarter by a group of 177 students who had taken the test twice. The mean gains ranged from a high of .6 of a standard deviation for the Social Sciences-History Test to a low of .3 of a standard deviation for the Mathematics Test. In relating the gains made on the GEs to grades in the courses corresponding to each test, different results were found for the five tests. Students with higher grades achieved greater gains on the Humanities, Natural Science, and Social Sciences-History Tests only. The authors conclude that "on the average the better students in the various courses come in those courses with better scores on the respective tests and show greater gains" (p. 5). French (1965) described mean gains on the five examinations for a group of 81 students. These gains are similar in pattern and magnitude to those reported by Harris and Booth. Kolb (1969) related gains to relevant course experiences for a sample

82 students tested twice. Significant gains were made by the students only on the English Composition and Natural Sciences Tests.

The score gains reported in the three foregoing studies do not necessarily indicate that a particular college has done a good job or a poor job. The GEs are designed to cover subject matter content as taught at different colleges with different curricula, methods, and materials. They do not necessarily reflect all the objectives and emphases of any one college. In addition, the lack of control groups makes it difficult to know whether the score gains were a result of instruction or simply a result of maturation or intellectual growth occurring within the first two years of college.

The relationship of the GEs' scores to amount of previous instruction in a subject generally provides support for the validity of the examinations as measures of academic achievement. A relationship, however, does not prove cause, and thus it cannot conclusively demonstrate that the scores are affected by instruction. Nevertheless, a lack of relationship between the GEs' scores and amount of previous instruction would have led one to question the validity of the tests.

Beanblossom (1969) correlated three GEs with the number of college credits taken in corresponding subjects. He concluded on the basis of his results that exposure to liberal arts courses "definitely" results in greater knowledge in natural sciences, "to some extent" in humanities, and "hardly at all" in social sciences and history. Selective factors, however, such as students taking more courses in their strong subjects, could account for these results.

The expectation that the tests' scores increase with the amount of formal college education completed has been confirmed by an analysis of the scores of 44,000 servicemen tested through the United States Armed Forces Institute. There appears to be a steady and significant progression of scores on all tests from those who have completed high school to those who have completed four years of college. Servicemen completing four years of college score about one standard deviation higher on each of the examinations than those who have not attended college.

The relationship of amount of high school preparation to the tests' scores was determined with the national freshman norming sample consisting of about 2500 second-term college students. Although the examinations were not intended to measure high school achievement,

scores on all tests correlated positively with the number of years of appropriate course work completed in high school.

Additional results relating to the construct validity of the examinations have emerged from the data collected with the national norming sample of approximately 2600 college sophomores. The scores of sophomores intending to major in different fields fell into expected patterns. The highest mean score on each of the five examinations was obtained by students intending to major in the field corresponding to the examination. For example, those intending to major in social sciences performed best on the Social Sciences-History Test while those majoring in humanities or fine arts scored highest on the Humanities Test.

The intercorrelations of the GEs indicate that to *some* extent all of the examinations except Mathematics are measuring the same ability or abilities—perhaps reading comprehension. The median intercorrelations found in five studies ranged from a low of .12 between Humanities and Mathematics to a high of .56 between English Composition and Humanities. It should be pointed out, however, that the intercorrelations are much lower than expected of reliable tests (above .9) measuring the same factors; thus, it is apparent that each test is also measuring some unique knowledge or skill.

Although the factorial composition of the GEs has not been determined, one could guess on the basis of the intercorrelations that two factors would account for most of the variance on the tests. The Mathematics Test would load high on a mathematical factor while the four other examinations would load high on a verbal factor.

Appropriateness of the tests for adults. One of the major target populations of the College-Level Examination Program consists of mature adults who have not had any formal education in college. The content of the GEs, however, is based on the program of study offered to freshmen and sophomores attending liberal arts colleges who are mostly in their late teens. Does the content or the format of the examinations place the older candidates at a disadvantage?

An analysis of the scores of approximately 44,000 servicemen on the GEs appears to suggest that the tests are no more difficult for the older than for the younger examinees. The oldest age group in this analysis, consisting of those of age 40 and over, was not the lowest scoring group on any of the examinations. In fact, this group had the highest mean score of any age group on the Social Sciences-History and Hu-

manities Tests. These two tests appear to be quite responsive to the accumulated value of life experience. The highest scores on the three other examinations occurred in the 22 to 24 age range. A limiting factor in the interpretation of this analysis is that the amount of formal education of servicemen at each age level was not known. While only 29 per cent of the sample had attended college, it is possible that the older age groups scored higher because they included more individuals with formal college education. Another possible explanation of the results is that the older servicemen in the sample were higher in ability or motivation as a result of self-selection.

French (1969) investigated the GEs' appropriateness with a sample of adult and black students. By using an inverse factor analysis on a matrix of the GEs' item responses he was able to identify 20 distinct hypothetical types of student, each defined by a certain set of items. Although the results suggest that the GEs do not give special advantage to any type of students, such as blacks or adults, it is difficult to have confidence in these results because the group of subjects used was small and unrepresentative.

Unfortunately, there have been no studies on the comparative validity of the GEs for different types of students. If the relationship between the tests' scores and a criterion is different for various groups of examinees, then the tests may not be equally appropriate for all groups. It may be, for example, that speed is a relatively more important factor for adults than for younger persons, and it might consequently invalidate the tests as measures of achievement for adults.

Conclusion. In general, the research summarized provides support for the validity of the GEs as measures of academic achievement. Many of the studies reviewed, however, do not lead to definitive conclusions. Results showing score gains after course exposure and positive relationships between the tests and amount of previous instruction have alternative interpretations. Correlations between the GEs and college grades obtained concurrently are moderately positive, but the validities of the tests for predicting success in upper-level studies are significantly lower than their validities for assessing current achievement level. The research methodology for validating the GEs can be improved by employing criteria other than grades, by using control groups in score-gain studies, and by partialing out contaminating factors in correlational studies. Nevertheless, the relationships found between the GEs and certain relevant variables provide ten-

tative support for the validity of the tests as measures of college-level achievement.

REFERENCES

- Beanblossom, G. F. The use of CLEP scores in evaluating liberal arts curriculum. Seattle: Bureau of Testing, University of Washington. Unpublished manuscript, 1969.
- Educational Testing Service. *ETS builds a test*. Princeton, N. J. 1965.
- French, J. W. Score gains on the Comprehensive College Tests during the first year. Sarasota, Florida: New College, Office of the College Examiner. Bulletin No. 10, 1965.
- French, J. W. Types of students defined by items in the CLEP General Series of Achievement Tests. Sarasota: New College. Unpublished manuscript, 1969.
- Harris, J. and Booth, E. An analysis of the performing of Georgia College freshmen and sophomores on the College-Level Examinations. Athens: Institute of Higher Education, University of Georgia. Unpublished manuscript, 1969.
- Koby, H. L. Report on the general education program at Rio Grande College as suggested by the College-Level Examination Program. Rio Grande, Ohio. Unpublished manuscript, 1969.

CONCURRENT VALIDITY OF A LITERATURE TEST IN RELATION TO SELECTION OF PERSONS FOR GRADUATE STUDY IN ENGLISH

JOSEPH P. SCHNITZEN AND JOHN A. COX

University of Houston

Problem. This study was performed to answer two questions: Is the field test of the Undergraduate Program (UP) in Literature (Educational Testing Service, 1969) a valid predictor of academic achievement in English as estimated by grade point average? If so, what level of test performance would be useful in selecting among applicants for graduate study in English?

Measurements. The instrument selected as a measure of academic achievement in literature was the field test, Literature, UP from Educational Testing Service (1969). It is a two-hour examination covering the following areas: poetry, fiction, drama, non-fiction, world literature, pre-Shakespeare, Shakespeare, English and American literature post-Shakespeare, and poetic metrics. It was judged to have adequate content validity. The test is recommended for evaluating academic achievement in literature at the college undergraduate level by the publisher.

Cumulative grade point average (GPA) was selected as one criterion of interest. This "score" was computed from grades in all courses attempted by a student except required physical education where a grade of A = 4 and F = 0. An English grade point average was also used as a criterion. It was computed using only grades from courses in English. For graduate students, only graduate work in English courses was used to compute a GPA.

Procedure. Three groups of students were of interest: seniors from the program preparing persons for teacher certification in English

(Eng. TE), seniors who were English majors (Eng.), and first-year graduate students in English (Grad.). The English department identified persons in the three groups. In Eng and Eng TE, 75 persons from each group were randomly selected, and letters were sent to each person strongly suggesting that the student come for testing at one of two specified times. Among graduates a similar technique was used, 50 letters being sent. No procedure was available to require that the student be tested.

Test administration was performed by personnel from the Counseling and Testing Service during May, 1970. In all 32 Eng TE, 35 Eng, and 33 Grad students were tested. Each of these persons was essentially a volunteer.

When testing was complete a search of student records was made in an attempt to find the grades made by each student for whom test scores were available. Grades were located for 31 Eng TE students, 33 Eng students, and 33 Grad students. In the Grad group two students who had completed fewer than three courses were eliminated from the correlational sample. Using the grade sheets, cumulative (Cum) GPA and Eng GPA were computed for Eng TE and Eng students. Only Eng GPA which included solely graduate English course grades, was computed for students in Grad.

From these data distributions, means and standard deviations for the Literature, UP test were computed for each group. These figures have been placed in Table 1. Correlation coefficients were computed for each GPA distribution and for the Literature test score distribution. Scatter diagrams were prepared and coefficients were computed from the diagrams. The results have been placed in Table 2.

Results. The figures in Table 1 show that there was about one-half a standard deviation difference between the mean literature scores of the Eng TE seniors and that of the Eng seniors, with the Eng TE mean being lower. This difference was statistically significant ($t = 12.31$). The observation is in line with expectation, since Eng TE students usually take about two fewer courses in English than do Eng students. Furthermore, the Grad students scored more than one standard deviation higher on the mean than did the Eng seniors, a difference that was also statistically significant ($t = 23.47$). Since members of the Grad group tended to be at the end of their first year of graduate study and to have completed an additional

TABLE 1
Literature Score Distributions by Group

Scores	Groups		
	Eng. TE	Eng.	Grad.
750-775		1	
725-749			1
700-724			3
675-699		1	3
650-674		1	6
625-649		2	7
600-624		1	5
575-599		1	1
550-574		2	
525-549	3	4	3
500-524	3	1	1
475-499	4	1	
450-474	4	8	1
425-449	9	4	
400-424	4	1	1
375-399	2	6	
350-374	1		1
325-349	2	1	
Mean	448	504	623
SD	52	91	78
N	32	35	33

year's course work above that of Eng group members, this finding was also expected. The data show that, while there was minor overlap in the three distributions, the Literature, UP test discriminated among the three groups of students in the expected direction and that the differences among means were rather large.

The correlation between Literature, UP test scores and Cum GPA (see Table 2) was .15 for Eng TE which is not statistically reliable and the corresponding correlation for the Eng group was .40, which is statistically significant at the .05 level. Thus, the test scores ex-

TABLE 2

Correlation Coefficients between Grade Point Averages and Literature Test Scores

	Literature Test Scores Group		
	TE Seniors	Eng. Seniors	Eng. Grad.
Cum. GPA	.15	.40*	—
Eng. GPA	.25	.48*	.32
N	31	33	31

* Significant beyond the .05 level.

hibited concurrent validity for overall academic performance as estimated by course grades only among English majors. When performance in English courses only was the criterion (Eng GPA), the relationships for test scores versus performance were somewhat higher. However, among Eng TE seniors the concurrent validity was not statistically reliable ($r = .25$). Neither was the concurrent validity among the Grad group ($r = .32$). Among the Eng group the relationship was significant beyond the .05 level ($r = .48$).

Conclusions. The Literature Test, UP has demonstrated evidence of construct validity in that scores from the test differentiated among groups of persons who had completed different amounts of course work in the field of English. The amount of differentiation was a practical amount. Among college senior English majors the Literature, UP test scores were moderately predictive of academic achievement. While the relationship between Literature, UP test scores and graduate academic achievement in English was not high enough to be statistically reliable in this study, the results were perhaps encouraging. In this study, the graduate students had already completed a year's work in graduate English courses. Thus they represented a selected sample. Each had applied for graduate school admission, had been evaluated on undergraduate performance and Graduate Record Examination performance, and had been accepted into graduate school before completing his course work. Had the group of senior English majors in this study been entered into graduate study of English without selection, the relationship between their test performance and graduate academic performance would have likely been higher than that found here for the Grad sample.

This study had shown that the Literature, UP test was valid in a construct sense and had concurrent validity among senior English majors. Thus, there was an answer to the first question to which the study was addressed. While no direct answer to the second question was given, the basis for making a judgement as to a specific cut-off score on Literature, UP test was furnished in the score distributions among the Eng and Grad groups.

Summary. Literature field test of the Undergraduate Program was administered to 32 English teacher education majors, 35 English majors, and 33 English graduate students. Test scores were found to be related to GPA among senior English majors. Mean scores for

each group were significantly different, both statistically and practically.

REFERENCE

- Educational Testing Service. *Handbook for Deans and Examiners. The Undergraduate Program for Counseling and Evaluation.* Princeton, New Jersey: Educational Testing Service, 1969.

PREDICTING QUALITY POINT AVERAGES IN MASTER'S DEGREE PROGRAMS IN EDUCATION

JERRY B. AYERS

Tennessee Technological University

THE prediction of success, as measured by quality point average, in Master's degree programs in education is a major concern of graduate schools. With increases in enrollment, there is a need to examine the quality point average prediction schemes that are in use by regional state universities. In studies conducted by Nunnery and Aldmon (1964), Owens and Roaden (1966), and Herbert (1967), the undergraduate quality point average (UQPA) was found to be the best predictor of graduate quality point average (GQPA). Miller (1970) summarized a number of predictive studies in which the Miller Analogies Test (MAT) has been used to predict GQPA. Herbert (1967) indicated a need for studies of the prediction of GQPA using a combination of the UQPA, MAT, and the National Teacher Examinations (NTE). Miller (1970) reported a correlation of .71 between the MAT scores and weighted total scores of the Common Examination of the NTE.

The inclusion of a measure of the ability of the graduate student to use the English language effectively in formal course work and in the preparation of papers, reports, and theses has been neglected in predictive studies of GQPA. It would appear that a test requiring use of the English language should be a significant predictor of success in a Masters program in education. An instrument such as the New Purdue Placement Test in English (PET) can sample a student's knowledge of "good English" (Wykoff, McKee and Remmers, 1955).

Purpose. The purpose of this study was to determine the relationship between each of several commonly used graduate school

admission criteria viewed as predictors and success in graduate school as measured by GQPA for graduates who had completed the Master of Arts in education at a regional state university.

Variables. The admissions criteria for full standing in a graduate program in education required a student to have completed his undergraduate program with a minimum UQPA of 2.50 (on a 4.00 scale) and to have completed the PET and MAT. In addition, a limited number of students presented scores on the NTE. For purposes of this study scores on the following parts of the NTE were used: Teaching Area Examination (TAE); Professional Education Scores (PES); Written English Expression (WEE); Social Studies, Literature, and Fine Arts (SLF); Science and Mathematics (SAM); General Education Subtotal (GES); Weighted Common Examination Total Score (WCE), and Composite NTE Score (CNS). The major criterion for success in graduate school was the completion of all requirements with a minimum GQPA of 3.00.

Sample. The sample was composed of those graduates ($N = 241$) who had completed the Master of Arts in education between June 1963 and August 1970 at a regional state university. All graduates had completed their programs of study with major emphasis in educational administration and supervision, curriculum and instruction, or guidance and counseling and had completed the MAT and PET. The NTE had been completed by 39 subjects.

Results and discussion. Intercorrelations, means and standard deviations for GQPA, UQPA, MAT and PET for each of the three groups and the total group of graduates are presented in Table 1. The UQPA and PET did appear to have a greater effectiveness in predicting GQPA, than did the MAT for the graduates of the administration and supervision and the curriculum and instruction programs. The findings with regard to UQPA were in agreement with the work of Nunnery and Aldmon (1964), Owens and Roaden (1966), and Herbert (1967). The correlation between PET and GQPA, for these two groups, was highly significant. This finding might be expected, since all graduates had completed a substantial number of courses in which their grades had been based largely on the preparation of reports and written projects. In addition all had completed a thesis or written project as part of their degree requirements. The correlation coefficients indicated that the criteria for grades in these programs were common with the performance

TABLE 1

*Intercorrelations, Means, and Standard Deviations for GQPA, UQPA, MAT, and PET for Graduates in Three Different Curricula and for the Total Group**

	UQPA	MAT	PET	Mean	SD
Administration and Supervision (<i>N</i> = 86)					
GQPA	.53	.41	.52	3.4	0.3
UQPA		.39	.64	2.6	0.4
MAT			.63	33.2	14.4
PET				132.4	26.5
Curriculum and Instruction (<i>N</i> = 47)					
GQPA	.69	.43	.61	3.7	0.3
UQPA		.36	.52	3.0	0.5
MAT			.60	38.0	14.5
PET				150.9	28.0
Guidance and Counseling (<i>N</i> = 108)					
GQPA	.28	.34	.34	3.6	0.3
UQPA		.27	.42	2.7	0.4
MAT			.66	38.2	15.2
PET				147.2	26.0
All Curricula (<i>N</i> = 241)					
GQPA	.49	.40	.50	3.5	0.3
UQPA		.34	.54	2.7	0.4
MAT			.65	36.4	14.9
PET				142.7	27.6

* All correlations are significant at or beyond the .01 level.

criteria required in earning grades at the undergraduate level and in performance on the PET.

The lack of substantial correlation between the criterion GQPA and each of the three predictors UQPA, MAT, and PET for graduates of the guidance and counseling program could be interpreted in part, since the curriculum involved performance criteria in laboratory course situations such as counseling techniques, test administration and interpretation, counseling interviews, and guidance intern experiences. Grades in courses of this nature reflected less emphasis on cognitive mastery of academic content and more on performance.

Intercorrelations of all variables for all curricula indicated that UQPA and PET were of about equal value in predicting GQPA. Correlations of MAT with GQPA for all groups and the total group were comparable to those previously reported (Miller, 1970). Multiple correlations, using the stepwise regression technique, for each of the groups were as follows: administration and supervision, .349; cur-

riculum and instruction, .563; guidance and counseling, .158; and all curricula, .332. The variables in this study served as the best predictors of success for the curriculum and instruction program in that 31.7 per cent of the variance was explained.

Table 2 presents the intercorrelations, means, and standard deviations for all variables and various subtests of the NTE for a limited group of graduates. An examination of the intercorrelations of the various measures indicated that the UQPA was the best overall predictor of GQPA. The correlation of the MAT with WCE was .66, which is in agreement with the value reported by Miller (1970). Using the stepwise regression technique, the equation for the prediction of GQPA was as follows:

$$\text{GQPA} = 1.770 + .285\text{UQPA} + .002\text{PET} \\ + .001\text{TAE} + .004\text{PES} - .005\text{WEE} - .009\text{SLF}$$

The multiple R for this equation was .554 which accounted for about 30.7 per cent of the variance in the criterion variable. The MAT did not enter into the equation.

Conclusions. The correlations between MAT and GQPA were typical of those found in the literature. They were interpreted as justifying the continued use of the MAT when only a single predictor is used. The validity coefficients presented in this study appeared also to justify the use of the UQPA and/or PET as predictors of GQPA in all curricula. The introduction of selected scores from the NTE when combined with UQPA and PET would seem to enhance the predictive qualities of the GQPA.

REFERENCES

- Herbert, D. J. A predictive study of quality point averages in graduate education courses. *The Journal of Educational Research*, 1967, 60, 218-220.
- Miller, W. S. *Manual for the Miller Analogies Test*, New York: Psychological Corporation, 1970.
- Nunnery, M. Y. and Aldmon, H. F. Undergraduate grades as indicators of success in master's degree programs in education. *Personnel and Guidance Journal*, 1964, 43, 280-286.
- Owens, T. R. and Roaden, A. L. Predicting academic success in master's degree programs in education. *The Journal of Educational Research*, 1966, 60, 124-126.
- Wykoff, G. S.; McKee, J. H. and Remmers, H. H. *Examiner's manual for The New Purdue Placement Test in English*, Boston: Houghton Mifflin, 1955.

ANOTHER CONTRIBUTION TO ESTIMATING SUCCESS IN GRADUATE SCHOOL: A SEARCH FOR SEX DIFFERENCES AND COMPARISON BETWEEN THREE DEGREE TYPES

DAVID A. PAYNE, ROBERT A. WELLS, AND ROBERT R. CLARKE
University of Georgia

THE literature of academic prediction in graduate education is filled with studies demonstrating correlations of various selection devices and grade point averages. The favorite devices are of course the Graduate Record Examination and the Miller Analogies Test. Results using these tests have certainly been unpredictable with reported validities ranging from .08 for the GRE-Verbal (Newman, 1968) to .47 for the GRE-Quantitative (Law, 1960). MAT validities ranging from .00 (Travers, 1948) to .69 (Gustad, 1950) have also been noted. Studies have yielded more supportive evidence for the MAT, with validities in the mid .20's and .30's. (e.g. Payne and Tuttle, 1966). Those concerned with selection in colleges of education will, however, find few readily available data on the potentially applicable and useful National Teacher Examinations.

To describe comparative validities between these three instruments was one of the purposes of the present investigation. In addition, since most predictive studies tend to focus on masters or doctoral students, or confound conclusions by combining these groups, prediction of success was concerned with a comparison of these two types of graduate degree programs. A few years ago the ultimate in degree acquisition was the masters. Now if one does not pursue a doctorate, one must at least secure a sixth year certificate. This alternative became the third degree type. And finally, a perusal of the literature reveals very little concern with sex differences in predictability of

success in graduate school. At other levels it is known that females are more predictable than males (Seashore, 1962).

Variables. Predictors involved in the present study were as follows:

1. Time to complete (TIME). Number of months from beginning to end of program or graduation. Doctoral students were measured from end of masters program.
2. National Teacher Examination—Common (NTE-C).
3. National Teacher Examination—Optional (NTE-O). One of 13 possible 80 minute examinations in various teaching areas.
4. Graduate Record Examination—Verbal (GRE-V).
5. Graduate Record Examination—Quantitative (GRE-Q).
6. Graduate Record Examination—Total (GRE-T). Simple addition of V and Q.
7. Miller Analogies Test (MAT). Raw scores used.
8. Undergraduate Grade Point Average (UA). Maximum average = 100.
9. Graduate Grade Point Average (GA).

Samples. All graduates of the College of Education during the academic year September, 1968 to August, 1969 composed the total group from which subgroups were determined. Theoretically data were available on a total group of 685 students, consisting of 314 males and 371 females. During this academic year 58 doctorates (both PhD and EdD), 503 master of education degrees, and 124 sixth year certificates were awarded. Applicants to the College of Education may submit scores on either the GRE, MAT, or NTE. Therefore, all students did not have scores on all tests. This fact may somewhat confound the interpretation of results. The subsample sizes are large enough in most cases to allow for a reasonable degree of confidence in the results. It was decided arbitrarily that any groups containing fewer than 10 subjects would not undergo correlational analysis.

Results and conclusions. Means, standard deviations, and correlations for the samples of Masters degree, sixth year certificate, and doctoral degree recipients are summarized in Tables 1, 2, and 3 respectively. Examination of Table 1 reveals that for the Masters people UA was about the best predictor of success. In addition it appears that females were somewhat more predictable than males. This last conclusion regarding greater female predictability was confirmed

TABLE 1
Descriptive Statistics Relating to Prediction of Success for Masters Degree Recipients

Predictor ^a	Males			Females			Total					
	N	\bar{X}	S	r^b	N	\bar{X}	S	r^b	N	\bar{X}	S	r^b
TIME	212	21.1	9.8	-.10	287	21.8	11.5	.00	499	21.5	10.8	-.04
NTE-C	179	601.7	75.2	.20	264	605.5	74.2	.23	443	604.0	74.6	.22
NTE-O	151	625.2	74.8	.15	243	634.7	74.3	.35*	394	631.1	74.6	.29
GRE-V	105	470.0	97.7	.36	112	489.6	101.9	.24	217	480.1	100.1	.29
GRE-Q	105	506.3	111.3	.04	112	453.7	110.1	.29*	217	479.1	113.5	.29
GRE-T	105	972.1	178.3	.21	112	943.2	181.5	.31	217	957.2	180.1	.26
MAT	29	49.9	15.4	.08	24	46.2	18.1	.07	53	48.2	16.6	.06
UA	194	80.7	3.7	.17	264	83.5	4.3	.33*	458	82.3	4.3	.31
GA	210	90.4	2.1		291	91.2	2.3		501	90.9	2.3	

^a Symbols for the variables are explained in the text.

^b Correlations with graduate grade average; N's for correlations average four fewer than those reported in first column for means.

* Correlation significantly higher for this sex ($p < .05$).

TABLE 2
Descriptive Statistics Relating to Prediction of Success for Sixth Year Certificate Recipients

Predictor ^a	Males			Females			Total		
	N	\bar{X}	S	r^b	N	\bar{X}	S	r^b	Total
TIME	42	22.6	12.8	.23 ^c	58	20.9	9.4	-.14	100
NTE-C	59	638.2	54.8	.24	66	630.2	55.9	.06	125
NTE-O	59	688.8	64.7	.20	65	715.7	52.5	.17	124
GRE-V	17	480.6	85.7	.17	16	468.6	93.1	.42	33
GRE-Q	17	502.9	106.5	.50	16	442.5	139.6	.14	33
GRE-T	17	983.5	143.1	.47	16	911.3	213.2	.27	33
MAT	4	53.3	11.5		2	46.5	12.0		6
UA	40	81.8	4.3	.00	60	83.8	4.4	.19	100
GA	58	91.0	1.9		66	91.7	1.8		124
									21.6
									634.0
									702.9
									474.9
									473.6
									948.5
									51.0
									83.0
									91.4
									10.9
									55.3
									60.0
									88.1
									125.5
									181.4
									11.0
									4.5
									1.9

^a Symbols for the variables are explained in the text.

^b Correlations with graduate grade average; N's for correlations average four fewer than those reported in first column for mean.

^c Correlation significantly higher for this sex ($p < .05$).

TABLE 3
Descriptive Statistics Relating to Prediction of Success for Doctorate Recipients

Predictor ^a	Males			Females			Total		
	N	X	S	r ^b	N	X	S	r ^b	S
TIME	43	30.1	11.2	.00	13	36.8	19.1	.02	31.7
NTE-C	30	649.9	69.9	.42	8	706.1	92.5		661.7
NTE-O	22	678.9	58.3	.28	7	782.9	51.2		704.0
GRE-V	40	513.5	95.1	.18	11	585.5	69.2	-.74 ^c	529.0
GRE-Q	40	528.0	107.1	.20	11	476.4	65.0	.65	516.9
GRE-T	40	1041.5	154.9	.24	11	1058.2	81.0	-.05	1045.1
MAT	12	45.6	4.5	.52	3	65.7	17.5		49.7
UA	39	83.4	3.7	.22	11	85.9	5.4	.32	84.0
GA	45	89.5	2.0		13	90.5	1.4		89.7
									1.9

^a Symbols for the variables are explained in the text.

^b Correlations with graduate grade average; N's for correlations average four fewer than those reported in first column for means.

^c Correlation significantly higher for this sex ($p < .05$).

when multiple correlations were computed from a combination of NTE-C, GRE-T, and UA. Multiple R 's of .34 ($N = 61$), .45 ($N = 66$), and .32 ($N = 127$) were found for males, females, and the total group, respectively. The multiple R 's still did not reflect a significant increase over the UA-GA zero order correlation.

In contrast to the Masters people, success in the Sixth Year Program (Table 2) can best be predicted from the GRE, with the Quantitative score being most effective for males, Verbal for females, and the Total Score for the combined group. Although no sex differences were noted in predictability, it was found that males tended to have higher average scores on most variables.

Success of the doctoral degree recipients (Table 3) was best predicted by the NTE-C. As with the Sixth Year people, no sex differences were noted, but perhaps a selection bias (self or institutional) was evidenced by higher average predictor scores for females. In some cases the sex differences were quite large. The large negative GRE-V correlation with GA was considered a result of sampling error, as might many of the correlations based on small samples.

Failure of the MAT to meet the writers' expectations might be somewhat a function of the lack of data. A sufficient number of cases to make a real test of predictive efficiency were not available.

In general it took females longer to complete the program than males. This time variable, however, was not an effective predictor.

It was almost impossible to determine a "most predictable" program. On either a comparative or absolute scale, differential predictability was almost a "toss up." When each program was associated with its best predictor, the results were about the same. A trend was noted such that UA, GRE, and NTE each in turn became best predictors of success in Masters, Sixth Year, and Doctoral programs. This might be interpreted as an increasing institutional and content emphasis on what constitutes success.

REFERENCES

- Gustad, J. W. Comparison between the Miller Analogies Test and the Graduate Record Examination as predictors of success in graduate training. Paper read at Midwest Psychological Association, 1950.
- Law, A. The prediction of ratings of students in a doctoral training program. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1960, 20, 847-851.

- Newman, R. I. GRE scores as predictors of GPA for Psychology graduate students. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 433-436.
- Payne, D. A. and Tuttle, Cynthia. The predictive relationship of the Miller Analogies Test to objective and subjective criteria of success in a graduate school of education. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1966, 26, 427-430.
- Seashore, H. G. Women are more predictable than men. *Journal of Counseling Psychology*, 1962, 9, 261-270.
- Travers, R. M. W. Unpublished study (1948). Reported in Miller, W. S. *Manual for the Miller Analogies Test*. New York: Psychological Corporation, 1960.

PREDICTION OF CHOICE OF AND SUCCESS IN AGRICULTURE AS A COLLEGE MAJOR¹

JAMES M. RICHARDS, JR.
American Institutes for Research

It is a cliché that mankind is in a desperate race between growth in population and growth in food supply. Moreover, there is considerable reason to believe that future increases in agricultural production, both in the United States and in other countries, will depend increasingly on the skill and level of training of agricultural workers rather than on purely technical improvements such as the introduction of hybrid seed varieties (Brown, 1967). For example, an important characteristic, compared to earlier seed varieties, of the seed varieties comprising the "Green Revolution" in underdeveloped countries is greater responsiveness to the application of fertilizer.

This implies that agricultural education at the college level will continue to be very important in the United States despite increasing urbanization and a decreasing proportion of the labor force engaged in farming. Welch's recent study (1970) suggests that a substantial part of the contribution of education to the farm earnings of college graduates should be attributed to increased ability to keep pace with improvements in available inputs and in the allocation of these inputs among various uses. These are just the characteristics likely to be required for future gains in agricultural production, and for maintaining a high level of productivity without destroying the ecosystem.

Accordingly, the purpose of this study was to investigate the

¹ Supported by funds from the U.S. Office of Education under grant number OEG-0-9-610065-1367. Opinions expressed are the author's.

validity of Project TALENT (Flanagan, Dailey, Shaycoft, Gorham, Orr, and Goldberg, 1962; Flanagan, Davis, Dailey, Shaycoft, Orr, Goldberg, and Neyman, 1964; Flanagan, Cooley, Lohnes, Schoenfeldt, Holdeman, Combs, and Becker, 1966) tests for predicting choice of and success in agriculture as a college major.² The overall goal of Project TALENT is to understand the nature and development of the talents of American young people. Because the questions involved are essentially developmental, the methodology has been, and continues to be, longitudinal. Specifically, in 1960 a probability sample was drawn of approximately 5 percent of the high schools in the United States. The 400,000 students in grades 9 through 12 attending the sampled high schools were administered two days of educational and psychological inventories specially constructed for Project TALENT. The student inventories included measures of general ability, specialized aptitudes, interests, personality, student activities, home background, and plans for the future. The overall design of Project TALENT calls for follow-up studies at intervals of one, five, ten, and twenty years after each class was graduated from high school. Thus Project TALENT provides the first long-range longitudinal study of a representative sample of students assessed with a comprehensive set of psychological, educational, and personal measures.

The present paper will summarize some of the results of the original assessment and the five-year follow-up. Although data collection for all four five-year follow-ups has been completed, because of the great mass of the Project TALENT data and the delays inherent in coding and keypunching responses to the follow-up questionnaires, not all of the merged files are complete at the present time (fall of 1970). Therefore, this paper is restricted to students who were enrolled in the eleventh and twelfth grade in 1960. The study is further restricted to men, because of the very small number of women who choose agriculture as a major.

Three criterion measures were used: (a) choice of agriculture as intended major on the original assessment, (b) designation of agri-

² Because of the importance of (and the author's specific interest in) agriculture, the present paper attempts to pull together data about agricultural majors in an accessible form. Some of the same data were also incorporated in a less integrated form in a more general treatment of college majors (Richards, 1970).

culture as actual major on the five-year follow-up, and (c) Grade Point Average in major field as self-reported on the five-year follow-up by those who designated agriculture as their actual major.

The predictors consisted of 31 scores, including three a priori composites of the TALENT ability tests (Verbal Composite, Quantitative Composite, and Technical Composite), the TALENT vocational interest test, the TALENT personality test, and a measure of socioeconomic status. These tests have been described in detail in earlier Project TALENT reports (Flanagan et al., 1962, 1964). The aptitude composites were chosen for similarity to commonly used aptitude composites and for independence from each other in the sense of having no sub-tests in common.

Table 1 shows the correlations between these predictors and each of the criteria for 11th and 12th graders separately. Correlations with choice are point biserials comparing students choosing agriculture with all other college students while correlations with GPA are Pearson product-moment correlations based only on those who majored in agriculture and reported their GPA. The beta weights and multiple correlations are summarized in Table 2. However, since, cross-validation data are not available, it should be mentioned that shrinkage in the multiple correlation values could be expected.

In interpreting these results, it must be remembered that there are ceiling effects on the point biserials as a consequence of the small proportion (2.5% to 2.9%) of students choosing agriculture as a major. When these effects are taken into account, it appears that choice of and success in agriculture can be predicted with roughly equivalent accuracy from the overall test battery used in this study.

The most outstanding characteristic of students choosing agriculture as major—both on the original assessment and on the five-year follow-up—is that they had relatively high scores on the Interest in Farming scale. For both eleventh and twelfth grade students, the scale having the highest correlation with GPA is Personal Maturity from the TALENT personality test. It is somewhat surprising that the aptitude composites were not more highly correlated with GPA.

Another trend in these data is that the predictors of choice and success are not very similar. This trend complicates the task of a

TABLE 1
Correlations between Predictors and Criteria

Predictor	Choosing Agriculture as a College Major While in High School		Majoring in Agriculture in College		GPA in Agriculture for Those Majoring in it	
	11th (<i>N</i> = 407)	12th (<i>N</i> = 411)	11th (<i>N</i> = 239)	12th (<i>N</i> = 252)	11th (<i>N</i> = 253)	12th (<i>N</i> = 240)
Verbal Composite	-.10	-.10	-.08	-.08	.14	.13
Quantitative Composite	-.10	-.09	-.06	-.07	.25	.13
Technical Aptitude Composite	-.04	-.04	-.01	-.01	.14	.08
Sociability	-.03	-.01	-.03	-.03	.13	.15
Social sensitivity	-.04	-.03	-.03	-.04	.17	.11
Impulsiveness	-.01	-.02	-.02	-.03	.14	-.07
Vigor	.02	.03	.02	.02	.16	.11
Calmness	-.01	.00	.00	-.02	.15	.16
Tidiness	-.03	-.02	-.03	-.04	.07	.17
Culture	-.03	-.02	-.03	-.04	.02	.09
Leadership	.00	.01	-.01	-.01	.21	.17
Self-confidence	-.03	-.03	-.02	-.05	.11	.08
Mature personality	.00	.00	.00	.00	.30	.23
Interest in:						
Physical science, engineering, math	-.10	-.07	-.07	-.04	.16	.08
Biological science and medicine	-.07	-.05	-.05	-.04	.13	.07
Public service	-.06	-.05	-.07	-.05	.09	.01
Literary-linguistic	-.09	-.10	-.09	-.09	.00	.10
Social service	-.03	-.03	-.04	-.03	.05	.10
Artistic	-.06	-.08	-.06	-.07	.01	.01
Musical	-.06	-.07	-.06	-.05	.01	.03
Sports	.00	.02	.00	.00	.15	.05
Hunting and fishing	.09	.09	.09	.08	.03	.04

TABLE 1—Continued

Predictor	Choosing Agriculture as a College Major While in High School	Majoring in Agriculture in College	GPA in Agriculture for Those Majoring in it
Business management	-.06	-.05	.04
Sales	-.04	-.04	-.01
Computation	-.08	-.06	.05
Office work	-.03	-.02	.02
Mechanical-technical	.03	.03	.11
Skilled trades	.08	.07	.02
Farming	.25	.20	.10
Labor	.06	.05	-.07
Socioeconomic status	-.03	-.04	.01

TABLE 2
Beta Weights and Multiple Correlations

Predictor	Choosing Agriculture as a College Major While in High School		Majoring in Agriculture in College		GPA in Agriculture for Those Majoring in it	
	11th	12th	11th	12th	11th	12th
Verbal Composite	-.0381	-.0500	-.0544	-.0326	.0203	.1404
Quantitative Composite	-.0309	-.0201	+.0058	-.0412	.1427	.0099
Technical Aptitude Composite	-.0002	.0091	.0283	.0327	.0502	-.0226
Sociability	-.0145	-.0021	-.0312	-.0076	.0728	.1419
Social sensitivity	-.0313	-.0398	-.0159	-.0370	.0444	-.0739
Impulsiveness	-.0016	-.0181	-.0147	-.0217	.0480	-.1142
Vigor	-.0033	.0028	.0173	.0222	.0041	-.0201
Calmness	.0193	.0296	.0161	.0095	-.0782	.0226
Tidiness	-.0295	-.0273	-.0389	-.0350	-.1441	.0522
Culture	.0184	.0291	.0274	.0168	-.1087	-.0566
Leadership	.0361	.0421	.0271	.0295	.0744	.1296
Self-confidence	-.0170	-.0203	-.0014	-.0372	.0388	-.0561
Mature personality	.0551	.0325	.0377	.0410	.3013	.1748
Interest in:						
Physical science, engineering, math	-.0598	-.0571	-.0563	-.0100	.0048	-.1373
Biological science and medicine	-.0256	-.0110	.0002	-.0033	.0591	.0771
Public service	.0122	.0254	-.0131	.0186	.0684	-.1437
Literary-linguistic	-.0222	.0298	-.0270	-.0587	-.1341	.2030
Social service	-.0094	-.0224	-.0294	-.0070	.0281	.0035
Artistic	-.0306	-.0586	-.0499	-.0568	.0122	-.0946
Musical	-.0092	-.0142	-.0080	.0135	.0187	-.0420
Sports	-.0267	-.0203	-.0214	-.0357	.0680	-.0321
Hunting and fishing	-.0342	-.0360	-.0061	-.0286	-.0530	-.0261
Business management	-.0471	-.0402	-.0026	-.0345	-.0425	-.0520

TABLE 2—Continued

Predictor	Choosing Agriculture as a College Major While in High School	Majoring in Agriculture in College	GPA in Agriculture for Those Majoring in it
Sales	.0060	.0011	— .0074
Computation	.0002	— .0115	— .0094
Office work	— .0293	.0005	— .1404
Mechanical-technical	— .0065	— .0231	.1617
Skilled trades	— .0076	.0159	— .0209
Farming	.3226	.2378	.0438
Labor	— .0320	— .0317	.0791
Socioeconomic status	— .0014	— .0224	— .0839
Multiple r	.32	.26	— .0325
			— .1569
			— .0009
			.39

counselor dealing with students considering agriculture as a college major or a career. It is possible, of course, that one or the other sets of predictors will be more similar to the predictors of success and satisfaction on the job in agriculture for these same students. The ten-year and twenty-year Project TALENT follow-ups should throw additional light on this question.

REFERENCES

- Brown, L. R. The world outlook for conventional agriculture. *Science*, 1967, 158, 604-611.
- Flanagan, J. C., Dailey, J. T., Shaycoft, M. F., Gorham, W. A., Orr, D. B., and Goldberg, I. *Design for a Study of American Youth*. Boston: Houghton-Mifflin, 1962.
- Flanagan, J. C., Davis, F. B., Dailey, J. T., Shaycoft, M. F., Orr, D. B., Goldberg, I., and Neyman, C. A., Jr. *The American High School Student*. (Technical report to the U. S. Office of Education, Cooperative Research Project No. 635), Washington, D. C., University of Pittsburgh, Project TALENT Office, 1964.
- Flanagan, J. C., Cooley, W. W., Lohnes, P. R., Schoenfeldt, L. F., Holdeman, R. W., Combs, J., and Becker, S. J. *Project TALENT One-Year Follow-up Studies*. (Final report to the U. S. Office of Education Cooperative Research Project No. 2333.) Pittsburgh: Project TALENT Office, University of Pittsburgh, 1966.
- Richards, J. M., Jr. Who studies what major in college? Paper presented at the American Psychological Association, Miami, 1970.
- Welch, F. Education in production. *Journal of Political Economy*, 1970, 78, 35-59.

USE OF THE ROTC QUALIFYING EXAMINATION FOR SELECTION OF STUDENTS TO ENROLL IN ADVANCED COURSES IN ROTC AS JUNIORS

THOMAS M. GOOLSBY, JR. AND DONALD A. WILLIAMSON
University of Georgia

THE Reserve Officer Training Corps Qualifying Examination (RQ) has been used since 1954 to screen college students into the advanced military course offerings beginning in the junior year. Antecedents to the 1954 examination date back to the years immediately following World War II.

Equivalent forms, RQ-8 and RQ-9, were implemented in the spring of 1966. RQ contains measures of verbal and mathematical ability.

The present study was designed to determine the extent to which RQ is useful for screening students into Advanced ROTC courses.

Procedures. The RQ (Form 9) was administered to students applying for Advanced ROTC in the fall of 1969 at a large southeastern university when they neared the end of their sophomore year of college. Scholastic Aptitude Test (SAT) scores, Freshman Military Grade Averages (F-Mil), Freshman Grade Point Averages (F-GPA), Sophomore Military Grade Averages (S-Mil), Sophomore Grade Point Averages (S-GPA), Junior Military Grade Averages (J-Mil), Junior Grade Point Averages (J-GPA), Cumulative Grade Point Averages (C-GPA), and certain subject area grade averages were obtained from student records.

Intercorrelations among all measures were obtained.

Results. The means and standard deviations for all variables are presented in Table 1. The estimated KR_{20} reliability for RQ-8 and RQ-9 is reported to be .92 for the norming sample ($N = 300$).

Table 2 presents the intercorrelations among subtests scores and

TABLE 1

Means and Standard Deviations for Certain Variables
(*N* = 77)

Variable	Mean	SD
ROTC Qualifying Examination-Verbal (RQ-V)	47.27	10.21
ROTC Qualifying Examination-Quantitative (RQ-Q)	32.16	7.16
ROTC Qualifying Examination-Total Score (RQ-T)	79.19	13.62
Scholastic Aptitude Test-Verbal (SAT-V)	509.87	68.22
Scholastic Aptitude Tests-Quantitative (SAT-Q)	542.44	80.50
Scholastic Aptitude Tests-Total (SAT-T)	1054.49	122.40
Freshman Military Grade Average (F-Mil)	3.46	.52
Sophomore Military Grade Average (S-Mil)	3.09	.57
Junior Military Grade Average (J-Mil)	3.00	.80
Freshman Grade Point Average (F-GPA)	2.72	.54
Sophomore Grade Point Average (S-GPA)	2.67	.56
Junior Grade Point Average (J-GPA)	2.61	.64
Cumulative Grade Point Average (C-GPA)	2.69	.48
English Grade Point Average (E-GPA)	2.24	.61
Math Grade Point Average (M-GPA)	2.50	.86
Science Grade Point Average (Sc-GPA)	2.19	.71
Social Science Grade Point Average (SS-GPA)	2.47	.64

TABLE 2

Intercorrelations among Subtests Scores and Total Tests Scores for RQ and SAT
(*N* = 77)^a

	RQ-Q	SAT-V	SAT-Q	SAT-T
RQ-V	15	69	38	61
RQ-Q		15	65	49
RQ-T		58	64	71
SAT-V			79	

^a Decimals omitted.

TABLE 3

Correlations of Subtests Scores and Total Tests Scores for RQ and SAT with Certain GPA's (*N* = 77)^a

	F-Mil	S-Mil	J-Mil	F-GPA	S-GPA	J-GPA	C-GPA
RQ-V	16	27*	13	17	12	03	12
RQ-Q	10	30*	19	21	12	07	22*
RQ-T	12	32*	15	16	13	04	12
SAT-V	09	17	05	06	08	10	05
SAT-Q	14	29*	19	20	12	07	15
SAT-T	15	28*	15	16	10	11	11

^a Significantly different from zero at the .05 level.

^a Decimals omitted.

TABLE 4

*Predictions of J-Mil and J-GPA by Certain Freshman and Sophomore Grade Averages (N = 77)**

	J-Mil	J-GPA
F-Mil	50	26
S-Mil	49	24
F-GPA	20	45
S-GPA	50	65

* Decimals omitted.

TABLE 5

*Correlations of Subtests Scores and Total Tests Scores for RQ and SAT with Certain Subject Area GPA's (N = 77)**

	E-GPA	M-GPA	Sc-GPA	SS-GPA
RQ-V	24*	08	05	23*
RQ-Q	19	28*	26*	16
RQ-T	23*	08	17	20
SAT-V	20	02	07	18
SAT-Q	23*	23*	15	20
SAT-T	24*	12	11	22*

* Significantly different from zero at the .05 level.

* Decimals omitted.

TABLE 6

*Certain Multiple Predictions of J-Mil and J-GPA (N = 77)**

RQ-T	& S-GPA	vs J-Mil	= 50
RQ-T	& F-Mil	vs J-Mil	= 50
SAT-T	& F-Mil	vs J-Mil	= 51
SAT-T	& S-GPA	vs J-Mil	= 51
F-Mil	& S-Mil	vs J-Mil	= 57
RQ-T	& SAT-T	vs J-Mil	= 14
RQ-T	& SAT-T	vs J-GPA	= 12
F-GPA	& S-GPA	vs J-GPA	= 18

* Decimals omitted.

total test scores for RQ and SAT. The correlation of RQ-V and RQ-Q (.15) is substantially different from the correlation between SAT-V and SAT-Q (.79). The correlation between RQ-T and SAT-T (.71) is somewhat lower than had been obtained earlier (.81). Even though the relationships of .71 to .81 are moderately high, they are mostly irrelevant to and are no reasonable justi-

fication for the adoption and use of RQ for the purposes outlined by the users.

The correlations of subtest scores and total tests scores for RQ and SAT with certain GPA's in Table 3 show very little and mostly no relationship at the .05 level of significance. The RQ shows no relationship to J-Mil or J-GPA at the .05 level.

Table 4 shows F-Mil, S-Mil, and S-GPA to be the best single predictors of J-Mil.

Again, the correlation of subtests scores and of total tests scores for RQ and SAT with certain Subject Area GPA's in Table 5 show very little and almost no relationship at the .05 level.

The multiple predictors of J-Mil and J-GPA in Table 6 are not materially different from the comparable single variable predictors presented earlier.

The data presented in this paper suggest that the degree of the relationships between academic ability measures investigated and college performance (grade point average) is low and questionable in its practical significance.

The data presented in this paper cast substantial doubt on the advisability of the use of RQ for the selection of even a reasonably large proportion of subjects from the population applying for admission to Advanced ROTC.

A NOTE ON THE PREDICTIVE VALIDITY OF THE COOPERATIVE ALGEBRA III

CLIFFORD B. TATHAM AND ELAINE J. TATHAM

Ottawa University
Ottawa, Kansas

MANY colleges use a variety of instruments as placement examinations. The object of these examinations is to aid in placing a student in the appropriate level of a particular course. Such placement testing is frequently done in the areas of English and Mathematics. While most of the instruments used are valid and reliable, many advisors of college freshmen are frequently confronted with the problem of encouraging, or discouraging, a student to enroll in a particular course. Thus, the purpose of this paper was to determine the predictive validity of the Cooperative Mathematics Test: Algebra III (Coop Algebra III) test at a small liberal arts college. The criterion was course grade. In addition, the reliability of the instrument was estimated since the suggestion is made that each institution determine its own reliability estimate (Educational Testing Service, 1964).

Method. During a period of three years beginning in the Fall semester, 1967, 362 students took the Coop Algebra III, Form A, as a part of the placement testing program at Ottawa University, Ottawa, Kansas. Each student had indicated a possible interest in taking either a mathematics or chemistry class. Complete data were available for 113 students who enrolled in and completed the first semester of a mathematics class which consisted of some algebra and an introduction to calculus and analytic geometry (M101).

Hoyt's method of estimating reliability employing analysis of variance was utilized (Hoyt, 1941). In addition, Saupe's formula

for estimating reliability was used so that the ease of Saupe's method might be employed in future studies (Saupe, 1961).

A discriminant analysis (Wert, Neidt, and Ahmann, 1954) was utilized to determine whether students who completed the M101 course successfully (grade of C or better) could be differentiated from those students who completed the M101 course unsuccessfully (grade of D or F).

Results. Using Hoyt's method of estimating reliability, a reliability coefficient $r = .8695$ and a standard error of measurement of 2.52 were obtained. These compare favorably with the published reliability of .84 and a standard error of measurement of 2.66 (Educational Testing Service, 1964).

Saupe's method of estimating reliability also yielded $r = .8695$.

The discriminant analysis indicated there was a significant difference between those students who successfully completed the M101 course and those students who failed to complete the M101 course successfully ($F_{1,112} = 15.36; p < .01$).

The analysis resulted in the equation: $v = .001239x$. A critical value of $v = .011979$ was obtained; thus the critical Coop Algebra raw score was approximately 9.

Summary. The results of this study indicate that the Coop Algebra III is a reliable instrument when used as a placement test. In addition, the test scores can be utilized by advisors to aid a student in deciding whether or not the student can successfully complete a mathematics course the content of which consists of some algebra and an introduction to calculus and analytic geometry.

REFERENCES

- Educational Testing Service. *Cooperative Mathematics Tests Handbook*. Princeton, N. J.: Educational Testing Service, 1964.
- Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, 6, 153-160.
- Saupe, J. L. Some useful estimates of the Kuder-Richardson formula number 20 reliability coefficient. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1961, 21, 63-71.
- Wert, J. E., Neidt, C. D., and Ahmann, J. S. *Statistical Methods in Educational and Psychological Research*. New York: Appleton-Century-Crofts, 1954.

STATISTICAL ANALYSIS OF THREE CRITICAL THINKING TESTS

JOHN FOLLMAN AND WILLIAM MILLER

University of South Florida

ELDON BURG

U. S. Army

Introduction. The purpose of this study was to investigate the psychometric characteristics of three critical thinking tests by determining their item difficulty and discrimination indices, reliability coefficients, item validities, and basic dimensions. A Test of Critical Thinking Form G (Form G) (American Council on Education, 1951); the Cornell Critical Thinking Test Form Z (Form Z) (Ennis, 1961); and the Watson-Glaser Critical Thinking Appraisal Form ZM (Form ZM) (Watson and Glaser, 1964) were the tests used.

The tests were administered to students in a junior level educational psychology course at Wisconsin State University, Oshkosh in May, 1967. The numbers of subjects for the different analyses ranged from 190 to 227. Form G has 52 items, Form Z has 52 items, and Form ZM has 100 items.

Results. Mean item discrimination indices were .34 for Form G, .23 for Form Z, and .18 for Form ZM.

Corrected split-half, and KR-20 total test reliability estimates were, respectively, .792 and .819 for Form G, .548 and .632 for Form Z, and .655 and .667 for Form ZM.

Corrected split-half reliability estimates were moderately high for two and high for seven Form G subtests, moderately high for two and low for five form Z subtests, and moderately high for all Form ZM subtests.

Mean point bi-serial correlations were .297 for Form G, .200 for Form Z, and .168 for Form ZM.

Form G had a higher proportion of high, significant phi coefficient inter-item correlations than either Form Z or Form ZM.

Factor analysis produced 19 factors for Form G, 22 for Form Z, and 39 for Form ZM. With some exceptions, particularly Form G subtests, items did not load on factors consistent with the test makers' a priori subtest groupings, for all three tests.

Recognition-of-assumptions items loaded fairly strongly within unrotated factors for all three tests.

Conclusions. It was concluded that Form G, Form Z, and Form ZM are reliable tests, that Form G is the most useful test, and that refinement should be undertaken for Form Z, and Form ZM.

REFERENCES

- American Council on Education, *Instructor's manual, Test of Critical Thinking, Form G*. Washington, D. C., 1951.
- Ennis, R. H. *Form Z information sheet*. Ithaca, New York: Cornell University, (Dec. 21), 1961.
- Watson, G. and Glaser, E. *Watson-Glaser critical thinking appraisal: Manual*. New York: Harcourt, Brace & World, Inc., 1964.

CROSS-VALIDATION OF THE ORLEANS-HANNA ALGEBRA PROGNOSIS TEST AND THE ORLEANS- HANNA GEOMETRY PROGNOSIS TEST

JOANNE M. LENKE AND HAROLD F. BLIGH

Harcourt Brace Jovanovich, Inc.

BERNARD H. KANE¹

Pleasantville High School
Pleasantville, New York

THE procedure of cross-validation is a necessary step in establishing the predictive power of a prognostic instrument in which weights are applied to the separate parts making up the total score. The technique involves the determination of the "best" weights with an initial sample and the subsequent verification of these weights with a second sample from the same population.

A total score on the Orleans-Hanna Algebra Prognosis Test and on the Orleans-Hanna Geometry Prognosis Test, is a composite of (a) four student-reported past course-grades, (b) student-predicted course-grade in algebra or geometry, and (c) number right of the Prognosis Test work-sample items. For both of the Orleans-Hanna tests, an a priori weight of two was assigned to each of the five course-grade variables and a weight of one was assigned to the test score representing the number of work-sample items answered correctly. Validation studies (Orleans and Hanna, 1968 a, b) were undertaken during the norming of the tests to determine the multiple regression weights for predicting each of the four criteria of success: mid-year and final course grades, and mid-year and end-of-year achievement test scores. The multiple correlations were then compared to the zero-order correlations obtained with the a priori

¹ Formerly of Yorktown High School, Yorktown Heights, New York.

weights. The results of these studies indicated that the weights, as initially assigned, could be applied in the final scoring process without significant loss in prediction. These weights were thus incorporated into the standard scoring procedure and were used in computing the total score and related normative data reported in the test manuals.

The purpose of the investigation reported here was to determine the appropriateness of the assigned weights in predicting similar criteria for new samples of algebra and geometry students.

Method. In June, 1968, 335 eighth-grade students took the Orleans-Hanna Algebra Prognosis Test and 331 ninth-grade algebra students took the Orleans-Hanna Geometry Prognosis Test. The school system cooperating in this study had not been included in the original validation sample. In January, 1969, the Mid-Year Algebra Test and the Mid-Year Geometry Test were given to those students who had enrolled in algebra or geometry. At the same time, algebra and geometry teachers, without access to test scores, reported mid-year grades for their students. In June, 1969, algebra students were given the Lankton First-Year Algebra Test and geometry students were given the Howell Geometry Test. Final mathematics grades were collected independently of the achievement test scores. Results of the prognosis tests were not released to school personnel until all criterion data had been collected.

Summary statistics, and zero-order and multiple correlation coefficients are presented in Tables 1 and 2 for the algebra and geometry cross-validation samples, respectively.

TABLE 1

Correlations of the Orleans-Hanna Algebra Prognosis Test with Four Criteria of Success

	Prognosis Test		Criteria			
	Total Sample	Sample Taking Algebra	Mid-Year Grade*	Mid-Year Test	Final Grade*	Final Test
<i>r</i>						
<i>R</i>			.72	.73	.70	.77
<i>N</i>			.73	.73	.73	.78
Mean	335	194	194	193	190	183
SD	67.6	74.0	2.6	24.6	2.6	30.5
	16.7	13.1	1.2	8.5	1.2	10.1

* A = 4, B = 3, C = 2, D = 1, E/F = 0.

TABLE 2

Correlations of the Orleans-Hanna Geometry Prognosis Test with Four Criteria of Success

	Prognosis Test		Criteria			
	Total Sample	Sample Taking Geometry	Mid-Year Grade*	Mid-Year Test	Final Grade*	Final Test
<i>r</i>			.64	.74	.60	.78
<i>R</i>			.67	.75	.66	.79
<i>N</i>	331	239	239	239	230	217
Mean	48.6	50.5	2.5	27.6	2.4	21.6
SD	12.5	12.2	1.2	6.7	1.1	8.4

* A = 4, B = 3, C = 2, D = 1, E/F = 0.

Results and conclusions. Examination of Table 1 reveals that with the assigned a priori weights, each of the zero-order correlations (*r*) between the Algebra Prognosis Test and each of the four criteria of success is sufficiently close to the corresponding multiple correlation (*R*) coefficient obtained when the predictor variables were optimally weighted, to make further refinement unnecessary (Bligh, Lenke, Hanna, 1969). Similarly, Table 2 indicates that very little predictive efficiency is lost when the assigned weights are applied in scoring the Geometry Prognosis Test. It appears, therefore, that the weights deemed appropriate for the original validation sample and recommended as part of the standard scoring procedure are equally appropriate for the cross-validation sample. One can further conclude that these weights will be equally appropriate for other samples of students within the same population.

REFERENCES

- Bligh, H. F., Lenke, J. M., and Hanna, G. S. The contribution of grades and work sample tests to the prediction of mid-year and end-of-year success in high school mathematics. Paper delivered at the 1969 Annual Meeting of the National Council on Measurement in Education, Los Angeles, California, February, 1969.
- Orleans, J. B. and Hanna, G. S. *Orleans-Hanna Algebra Prognosis Test Manual*. New York: Harcourt Brace Jovanovich, Inc. 1968. (a)
- Orleans, J. B. and Hanna, G. S. *Orleans-Hanna Geometry Prognosis Test Manual*. New York: Harcourt Brace Jovanovich, Inc., 1968. (b)

A COMPARISON OF THE D-48 TEST AND THE OTIS QUICK SCORE FOR HIGH SCHOOL DROPOUTS

BRAD S. CHISSOM AND RALPH LIGHTSEY

Georgia Southern College

THE D-48 (Dominoes) Test has been used in its present form for a number of subject populations. The test, consisting of a series of nonverbal problems that use dominoes as the item format, is described as a nonverbal analogies test measuring general intelligence (Black, 1961). In a study of college students, Boyd and Ward (1967) obtained a correlation of .57 between the D-48 and the Otis Quick Score (Gamma Form FM).

The internal consistency (KR_{20}) reliability for the D-48 test was reported in the same study as .85. Additional studies across age levels and cultures revealed the rank-order of the item difficulties to be similar. Rafi (1967), Gough and Domino (1963), and Welsh (1967) all cited similar rank orders of the item difficulties despite differences in the age and cultural background of the subjects. Using a group of non-college French males, Pasquay and Doutrepoint (1956) reported results on the D-48 with subjects similar to the ones tested in this study.

Purpose and method. This study was designed to examine two measures of intelligence, the D-48 and the Otis Quick Score (Otis, 1954), when used with a population of male high school dropouts. The subjects were full-time enlisted personnel enrolled in a General Educational Development course conducted by the United States Army. Sixty-one male subjects, whose permanent homes represented 28 states, were included in the study. The two tests, the D-48 and the Otis Quick Score, were administered on two successive days during the regular classroom instructional period. The specific objectives of the study were: (a) to obtain a correlation between D-48

scores and scores on the Otis as an estimate of concurrent validity, (b) to compare the rank-order of the item difficulties obtained with the rank-orders reported in other studies using different subject populations, and (c) to make a comparison of the mean score obtained for the D-48 with the mean score obtained from a comparable group of subjects with a different cultural background.

Results and discussion. Reliabilities for the D-48 and the Otis were calculated by the odd-even, split-half method increased by the Spearman-Brown Formula. The reliabilities were .92 for the D-48 Test, and .88 for the Otis. The KR_{20} reliability was .85 for the D-48.

The correlation between the D-48 and the Otis was .27. Comparing this to the correlation of .57 reported by Boyd and Ward (1967) in their study with college students of the same age, the verbal ability called for in the Otis probably accounts for the decrease. The decrease in the relationship between the tests would seem to support the idea that the D-48 assesses nonverbal intelligence.

Comparisons of the rank-orders of the item difficulties were made between the rank-order obtained from the results of this study, the Welsh (1967) data for gifted high school students, and the Gough and Domino (1963) data for fifth and sixth grade pupils. The comparisons were made using the Spearman rank-difference correlation, and the data are shown in Table 1. The results agree with those reported by Rafi (1967), in which the correlation between the Gough and Domino difficulty ranks and difficulty ranks obtained from a sample of Lebanese men was .95. This evidence of the magnitude of the relationship between the item difficulty ranks indicates the comparability of the D-48 across age levels and cultural backgrounds.

Finally, the D-48 mean score of 19.03 for the 61 dropouts (see Table 2) was not significantly different from a mean of 19.78 for a sample of non-college French males, ages 20-25 (Pasquay and

TABLE 1
Intercorrelations for Three Item Difficulty Rank-Orders

Source of Item Ranks	1	2	3
1. High School Dropouts (Chissom)	—	.95	.91
2. High School Gifted (Welsh)		—	.94
3. Fifth and Sixth Graders (Gough-Domino)			—

TABLE 2

Means and Standard Deviations of Age, Grade Completed, D-48 Raw Scores And Otis Quick Score Raw Scores
(*N* = 61)

Variable	Mean	SD
Age	21.62	3.82
Grade Completed	9.51	1.48
D-48 Raw Score	19.03	6.33
Otis Raw Score	29.43	8.75

Doutrepoint, 1956). This result is additional evidence of the cross-cultural nature of the D-48 Test.

Summary. The D-48 and Otis Quick Score tests designed as measures of intelligence were administered to a group of male high school dropouts. Results indicated that the relationship between the two tests is less for an individual representing on the average between a ninth- and tenth-grade level than for college students. Further results indicated that the relative levels of D-48 test item difficulties are comparable across age levels, and for subjects with diverse cultural backgrounds.

REFERENCES

- Black, J. D. *The D-48 Test: Preliminary manual*. Palo Alto, California: Consulting Psychologists Press, 1961.
- Boyd, M. E. and Ward, G. Validities of the D-48 Test for use with College Students. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 1137-38.
- Gough, H. G. and Domino, G. The D-48 Test as a measure of general ability among grade school children. *Journal of Consulting Psychology*, 1963, 27, 344-49.
- Otis, A. S. *Otis Quick-Scoring Mental Ability Tests, Gamma Form AM*. New York: World Book Co., 1954.
- Pasquay, R. and Doutrepoint, G. Le Test des Dominoes (D-48). *Bulletin De Psychologie Scolaire Et D'orientation*, 1965, 5, 20-34.
- Rafi, A. A. The Progressive Matrices (1938) and the Dominoes (D-48) Tests: A cross cultural study. *British Journal of Educational Psychology*, 1967, 1, 117-19.
- Welsh, G. S. *Performance analysis of gifted adolescents on two intelligence tests*. Final Report, Contract U. S. Dept. HEW Project No. 7-C-009, Chapel Hill: University of North Carolina, 1967.

THE RELATIVE PREDICTIVE AND CONSTRUCT
VALIDITIES OF THE OTIS-LENNON MENTAL ABILITY
TEST, THE LORGE-THORNDIKE INTELLIGENCE
TEST, AND THE METROPOLITAN READINESS TEST
IN GRADES TWO AND FOUR: A SERIES OF
MULTIVARIATE ANALYSES¹

BARTON B. PROGER, JOHN R. MCGOWAN
ROBERT J. BAYUK, JR., AND LESTER MANN²

Research and Information Services for Education, King of Prussia, Pa.

RUTH L. TREVORROW AND EDWARD MASSA³

School District of Cheltenham Township, Elkins Park, Pa.

Introduction. The Otis-Lennon Mental Ability Test (Otis and Lennon, 1967a, 1967b) could be a major breakthrough in the field of measuring academic potential. No doubt a great deal of ongoing and future educational and psychological research will make use of the improved Otis instrument. However, as Groteleuschen (1969) pointed out, validity studies with the new test are scarce. The

¹ This study was supported cooperatively by RISE, an Elementary and Secondary Education Act of 1965 Title III project (OEG-1-67-3010-2696), the School District of Cheltenham Township, and a predoctoral USOE fellowship in educational research granted to J. R. McGowan at Lehigh University, Bethlehem, Pa. However, the opinions expressed herein do not necessarily reflect the policies or positions of the cooperating agencies. The senior author assumes responsibility for the methodological views, design, and analyses represented in this paper.

² BBP is now with Pennsylvania Resource and Information Center for Special Education, 443 South Gulph Road, King of Prussia, Pa. 19406. JRMCG is now with Southern Connecticut State College, Schwartz Hall, Administration Office, Room 105, 501 Crescent Street, New Haven, Conn. 06515. RJB is now with Research Division, School District of Philadelphia, Computer-Assisted Instruction Project. LM is now with National Pennsylvania Regional Resources Center, King of Prussia, Pa.

³ This paper is a revised, extended version of a talk delivered at the 1970 Annual Convention of the National Council on Measurement in Education

present study not only provides some initial information on the predictive validity of the 1967 Otis-Lennon test relative to the Lorge-Thorndike instrument, but also yields basic data on the factor analytic structure of the battery of tests. The relative predictive and construct validities of the Otis-Lennon, Lorge-Thorndike, and Metropolitan Readiness tests were investigated by several multivariate analyses: canonical correlation, factor analysis, and stepwise regression.

Procedure. The pupils selected for this study comprised the total enrollments of the second and fourth grades for the academic year 1968-1969 of a large suburban public school district in the Greater Philadelphia Area. The district is distinctly of upper socio-economic class. The majority of pupils are college-bound. Descriptive data (including the intercorrelations of all variables at second and fourth grades) for these pupils can be found in Table 1.

The second grade pupils were administered the Lorge-Thorndike Intelligence Test (L-T IT) (Level A, Form 2) in October, 1968, and the Otis-Lennon Mental Ability Test (O-L MAT) (Elementary I Level, Form J) in November, 1968. The Metropolitan Readiness Test (MRT) (Form A) had been given to 322 of the current group of 386 second grade pupils in April, 1967 (the end of their kindergarten enrollment).

The fourth grade pupils were administered the L-T IT (Level B, Form 1) in October, 1968, and the O-L MAT (Elementary II Level, Form J) in November, 1968. MRT (Form A) had been given to 316 of the 469 present fourth grade pupils in April, 1965 (the end of their kindergarten enrollment).

The main set of academic achievement criteria for the second grade pupils was the Stanford Achievement Test (SAT) (Primary II Battery, Form W), and for the fourth grade pupils was the SAT (Intermediate I Battery, Form X). Both tests were given in April, 1969. Besides using standardized SAT criteria, it was considered feasible and interesting to have the teachers of both second and fourth graders rate their pupils on academic ability relative to the

in Minneapolis, Minn., March 4, 1970 (Session 1, Research on Testing). The authors wish to thank the personnel of Cheltenham, especially Dr. Lawrence Green and Mr. Philip Butler. Also, deep appreciation is expressed to Mr. David March, Predoctoral Educational Research Fellow at Lehigh, for his assistance in using the Control Data Corporation 6400 computer at Lehigh's Computer Center.

TABLE 1
Intercorrelation Matrix and Descriptive Information*

	O-L MAT Tot	L-T IT Nonv	L-T IT Tot	MRT	TR Rdg Comp	TR Arit Comp	TR Arit App	SAT Word Mean	SAT Para Mean	SAT Sci SS	SAT Spel	SAT Word SS	SAT Lang	SAT Arit Comp	SAT Arit Conc	SAT Arit App
1	83			62	55	59	56	61	53	38	53	53	59	44	68	
2	72	70														
3																
4																
5	51	48		48	52	51		53	45	36	42	42	48	32	50	
6	71	56		45	59	53		60	44	48	51	51	57	33	55	
7	50	53		37	65	78		77	46	67	62	62	70	51	62	
8	58	61		42	76	85		64	44	60	61	61	66	58	67	
9	60	60		42	75	83										
10	73	78		46	72	44	90	85	52	70	66	66	76	44	61	
11	75	75		44	75	49	59	82	55	32	43	43	48	52	68	
12																
13	61	64		33	62	48	54	68	67	61	73	69	68	57	59	
14	66	61		41	64	49	55	66	65	70	46	52	60	52	66	
15	75	73		48	74	56	56	72	78	52	60	65	70	68	61	
16	49	45		31	53	56	57	53	53	52	65	65	70	68	61	
17	68	64		48	64	60	64	66	68	52	65	65	70	68	61	
18	66	63		38	62	54	60	62	67	48	56	56	63	64	80	
SD ₂	14.89			11.61	13.37	1.15		7.15	11.06	5.76	7.66	11.95	10.09	6.69	8.51	
\bar{X}_2	114.54			113.34	64.36	3.57		20.95	35.96	21.52	16.14	44.49	42.25	23.96	24.01	
SD ₄	14.40	13.76	14.35	12.07	1.04	1.02	1.08	7.08	9.77	9.32	8.77	8.77	14.13	7.01	5.79	5.96
\bar{X}_4	120.37	113.67	118.58	78.30	3.47	3.50	3.30	26.16	39.76	35.96	47.90	47.90	87.76	24.05	22.28	20.75

* Second-grade intercorrelations are presented above the diagonal, while fourth-grade intercorrelations are given below the diagonal. All intercorrelations are rounded to two decimal places, with decimal points omitted. Blank rows or columns indicate that the variable was not used at that grade level. $N_2 = 322$ and $N_4 = 316$.

school district as a whole (second or fourth) in specific areas. In each academic area selected, the rating was accomplished by a five-choice continuum. The code was: 5, outstanding; 4, good; 3, average; 2, below average; and 1, very limited. For second grade, two teacher criteria employed were termed "Reading Comprehension," and "Arithmetic Computation." For fourth grade, four teacher criteria used were termed "Reading Comprehension," "Arithmetic Computation," "Arithmetic Concepts," and "Arithmetic Application." The use of teacher rating (TR) in this study is similar to that of Kim, Anderson, and Bashaw (1968) in their canonical correlation study.

Analyses. Several sets of analyses were undertaken: canonical correlation, principal-components factor analysis, and stepwise regression. The BMD06M canonical correlation technique (Dixon, 1967, pp. 207-214) was selected to investigate the overall patterns of significant relationship and to suggest the factor analytic structure of the total battery of predictors and criteria. However, the detailed factor analysis was left for the BMD03M principal-components factor analysis (Dixon, 1967, pp. 169-184). Once the factorial composition of tests had been established, the final stage of the validity analyses was to assess the relative predictive powers of the three tests in question by the BMD02R stepwise regression program (Dixon, 1967, pp. 233-257d.).

The authors were particularly interested, from a methodological viewpoint, in how the canonical correlational analysis compares with principal-components factor analysis in the factorial-structure sense.

Many different philosophies underlie the interpretation of canonical correlation. Duntzman and Bailey (1967) discussed the differences between factor analysis and canonical correlation, while Ohnmacht and Olson (1968) emphasized the similarities. Maxwell (1961) interpreted canonical vector weights and compared them with factor analytic weights. By applying both principal-components factor analysis and canonical correlation to the same set of data, the authors hoped to assess the similarities and differences in the resulting factor-analytic structures.

Results—Construct validity. Table 2 presents the canonical correlation vector weights for second and fourth grades. Bartlett's Test

TABLE 2
Canonical Correlation Relationships

	Second Grade			Fourth Grade				
	.81	.27	.14	.88	.45	.25	.19	
Predictors	O-L MAT Tot	.57	1.18	.36	.39	-.07	1.90	-.42
	L-T IT Verb				.47	-1.12	-1.28	.60
	L-T IT Nonv				.13	1.36	-.35	.57
	L-T IT Tot	.20	-.45	-1.21				
	MRT	.41	-.92	.70		.05		-1.06
Criteria	TR Rdg Comp	.29	-.06	.21	.08	-.69	.72	-.90
	TR Arit Comp	.08	-.40		.04	.01	-.24	.14
	TR Arit Conc			-1.16	-.02	.51	-.98	.40
	TR Arit App				.13	.11	-.15	.07
	SAT Word Mean	-.01	-.14	-.16	.30	-.92	-.93	-.24
	SAT Para Mean	.22	-.25	-.18	.15	.24	.42	.70
	SAT Sci SS	.13	.04	-.65				
	SAT Spel	-.20	-.90	.45	.02	-.52	.06	.61
	SAT Word St S	.18	.16	.42	.05	.50	.57	.08
	SAT Lang	.17	.21	.39	.26	.21	.03	-.64
	SAT Arit Comp	-.11	.52	-.26	-.14	-.16	.29	-.04
	SAT Arit Conc	.38	.72	.82	.14	.56	-.36	-1.09
	SAT Arit App				.12	.22	.60	1.13

Note—Bartlett's Test of Wilks' Lambda Criterion shows that for second grade, the first canonical relationship is significant at the .01 level, while for fourth grade, the first two canonical relationships are significant at the .005 level.

of Wilks' Lambda Criterion (Cooley and Lohnes, 1962, p. 37) was applied.

Findings for fourth grade pupils. Out of the four canonical relationships given for fourth grade, only the first two are significant at the .05 level. The first relationship ($R_c = .88$), or factor, that underlies both predictor and criterion variables apparently emphasizes the verbal composition of both the O-L MAT and L-T IT (verbal) tests in terms of dimensions common to the SAT: Word Meaning and to a lesser extent, SAT: Language. All these variables have positive weights. Thus, the first canonical variate is fairly easily interpreted. The second variate ($R_c = .45$) deals only with the verbal and nonverbal components of the L-T IT as predictors; in other words, the construct validity of the intelligence scores of the L-T IT may be somewhat clarified. One sees that the nonverbal component is most highly related to TR:Arithmetic Concepts, SAT:Word Study Skills, and SAT:Arithmetic Concepts. One should note, however, that the SAT:Word Study Skills does appear to be out of place with respect to the part of the dichotomy in this canonical variate that is presumably "nonverbal." All the nonverbal variables

have positive weights. The verbal component of the second canonical variate consists of the negatively weighted variables of L-T IT Verbal, TR:Reading Comprehension, SAT:Word Meaning, and SAT:Spelling.

Next, the canonical variate analysis involving the two significant dimensions for fourth grade pupils is compared with the corresponding principal-components analysis also yielding two factors. The latter results are given in Table 3. When only eigenvalues greater than one were used, two factors were extracted. The first factor shows that, in the total set of variables, the O-L MAT, L-T IT, and MRT are loaded heavily on the standardized SAT verbal criteria; this result is in accord with the first canonical variate above. However, the second factor points out one of the basic differences between canonical correlation and factor analysis, as discussed by Duntzman and Bailey (1967). In Table 3, one sees that the second factor loads primarily on numerical criteria, and only modestly on the predictor tests in question. Thus, while canonical correlation "forced" the predictor tests to have high weights on each variate, the principal-components factor analysis allowed the numerical variables to cluster together ("internal factor analysis") without the inclusion of the predictor tests.

Findings for second grade pupils. Turning to second grade, one sees the canonical variate weights in Table 2. The first variate ($R_c = .81$) emphasizes the loading of the O-L MAT and MRT on a dimension associated with the SAT:Arithmetic Concepts, SAT:Paragraph Meaning, and TR:Reading Comprehension criteria. If one is willing to believe that arithmetic concepts require verbal as well as numerical abilities, then the first variate can be termed a verbal factor underlying both predictors and criteria. In the second canonical variate ($R_c = .27$), the O-L MAT is weighted positively along with the SAT:Arithmetic Computation and SAT:Arithmetic Concepts, while the MRT and, to a lesser extent, the L-T IT (total scores) are weighted negatively in association with SAT:Spelling and TR:Arithmetic Computation. Using the relative weights on the criteria in the second variate, one sees that the O-L MAT is more closely related to the numerical criteria than to the verbal criterion, while the opposite is true for both the MRT and L-T IT. However, one must use caution in interpreting the

TABLE 3
Principal-Components Factor Analyses*

Variable	Second Grade			Fourth Grade		
	Loading on Factor A	Loading on Factor B	Communality	Loading on Factor A	Loading on Factor B	Communality
MAT	33	79	73	83	32	80
IT Verb				81	33	77
IT Nonv				64	44	59
IT Tot	22	76	63			
T	44	59	54	51	27	33
Edg Comp	73	46	75	64	58	74
Arit Comp	70	45	69	23	90	86
Arit Conc				37	86	88
Arit App				37	86	88
Word Mean	74	46	75	87	23	80
Para Mean	71	52	78	84	31	80
Sci SS	20	75	61			
Spel	87	13	77	70	30	58
Word St S	78	30	70	71	35	63
Lang	74	44	74	78	43	79
Arit Comp	67	21	50	39	62	54
Arit Conc	53	64	70	64	55	71
Arit App				61	52	64

All loadings are rounded to two decimal places, with decimal points omitted. Blank rows indicate that the variable was not used at that grade level.

second canonical variate, since this result was not significant at the .01 level.

In Table 3, one sees the principal-components analysis for the total battery of tests in second grade. The first factor is loaded mostly with verbal criteria, with slightly lower loading for the two numerical criteria; the predictor tests have relatively low loadings. In the second factor, the O-L MAT and L-T IT load most heavily on SAT:Science and Social Studies. The factor-analytic picture is not so clearly defined for the second grade as for the fourth grade sample.

Results—Predictive validity. Several separate stepwise regression analyses were conducted in selected verbal and numerical areas. Table 4 presents the standardized regression equation results.

Prediction of SAT achievement measures. In general, at the second-grade level, the O-L MAT appeared to be a better predictor of standardized verbal and numerical achievement test performance than was the L-T IT. However, despite the large proportion of variance in any one particular criterion measure explained by the

TABLE 4
Stepwise Multiple Regression Analyses for Predicting Selected Stanford and Teacher Rating Variables

Grade	Variable To Be Predicted	Stanford				Teacher Rating					
		Variable Entered	R*	R ² *	Inc In R ² *	F To Ent	Variable Entered	R*	R ² *	Inc In R ² *	F To Ent
Second	Para Mean (Rdg Comp)	O-L MAT Tot	61	38	38	191.70	MRT	59	35	35	170.81
		MRT	69	48	10	61.14	O-L MAT Tot	67	45	10	57.11
		L-T IT Tot	70	49	02	9.78	L-T IT Tot	68	46	02	10.89
	Arit Comp	O-L MAT Tot	44	19	19	76.73	O-L MAT Tot	56	32	32	147.03
		MRT	45	20	01	4.17	MRT	62	39	07	37.92
		L-T IT Tot	45	21	00	.76	L-T IT Tot	64	41	02	13.27
Fourth	Para Mean (Rdg Comp)	O-L MAT Tot	75	57	57	410.99	O-L MAT Tot	71	50	50	315.86
		L-T IT Verb	78	62	05	39.28	L-T IT Verb	74	55	04	30.54
		MRT	79	62	00	2.21	MRT	74	56	01	6.72
	Arit Comp	L-T IT Nonv	79	62	00	.96	L-T IT Nonv	74	56	00	.01
		O-L MAT Tot	49	24	24	100.09	L-T IT Nonv	53	28	28	123.81
		L-T IT Nonv	51	26	02	7.94	L-T IT Verb	57	33	05	21.29
	Arit Conc	L-T IT Verb	51	26	00	.96	MRT	58	34	01	3.86
		MRT	51	26	00	.67	O-L MAT Tot	58	34	00	.22
		O-L MAT Tot	68	47	47	274.04	L-T IT Verb	61	38	38	188.15
	Arit App	L-T IT Nonv	72	52	06	36.55	L-T IT Nonv	66	43	06	32.97
		MRT	73	53	01	8.01	MRT	67	44	01	6.16
		L-T IT Verb	74	54	01	4.03	O-L MAT Tot	67	45	00	.49
		O-L MAT Tot	66	44	44	245.83	L-T IT Verb	62	38	38	191.80
		L-T IT Nonv	69	48	04	25.74	L-T IT Nonv	66	44	06	31.31
		L-T IT Verb	70	49	01	4.47	MRT	67	45	01	5.92
		MRT	70	49	00	.08	O-L MAT Tot	67	45	00	1.84

* Entries are rounded to two decimal places, with decimal points omitted.

O-L MAT alone, the MRT usually accounted for a roughly similar amount. At the fourth-grade level, the regression analyses become somewhat more refined because of the distinction in the L-T IT of verbal and nonverbal intelligence. For the SAT criteria, again the O-L MAT appeared to be a slightly more valid predictor than the L-T IT, although the differences were not nearly so marked as in second grade.

Specifically, for the SAT Paragraph Meaning Test, the best single predictor was the O-L MAT in both second and fourth grades ($R_2 = .61$ and $R_4 = .75$).⁴ For the SAT Arithmetic Computation Test, again the best single predictor was the O-L MAT in both grades ($R_2 = .44$ and $R_4 = .49$). In fourth grade, the O-L MAT was the best single predictor of the SAT Arithmetic Concepts Test ($R_4 = .68$) and of the SAT Arithmetic Applications Test ($R_4 = .66$).

Prediction of TR achievement measures. Also of interest to the predictive validity aspect of the study was the relationship of the five-point teacher ratings in selected verbal and numerical areas to the three predictor tests. The best single predictor of TR:Reading Comprehension was the MRT in second grade ($R_2 = .59$) and the O-L MAT in fourth grade ($R_4 = .71$). For TR:Arithmetic Computation, the best single predictor in second grade was the O-L Mat ($R_2 = .56$) and in fourth grade was the L-T IT Nonverbal test ($R_4 = .53$). Finally, the L-T IT Verbal test was the best single predictor for fourth-grade TR:Arithmetic Concepts ($R_4 = .61$) and for fourth-grade TR:Arithmetic Applications ($R_4 = .62$). Clearly, in the case of the 5-point teacher ratings, the O-L MAT does not have a virtual monopoly on the highest predictive validities. Indeed, at the fourth-grade level, the L-T IT Verbal and Nonverbal tests seem to be superior to the O-L MAT.

Just how much confidence can be placed in the teacher ratings can be partially answered by their relative canonical variate weights in Table 2. The high teacher rating canonical vector weights might be reflecting at the elementary school level the oft-repeated finding that teacher-given grades are more valid predictors of later achievement than are standardized tests.

⁴ All predictive validities are given in terms of the multiple correlation coefficient at the first stage of the stepwise model building; that is, the correlation involves only the best single predictor and the criterion.

Summary. Canonical correlation and principal-components factor analysis were employed to study the factorial construct validity of the O-L MAT, L-T IT, and MRT. The results demonstrated some similarities and differences between the two analytical approaches to construct validation.

Further, stepwise multiple regression was used to establish the relative predictive validities of the three tests in selected verbal and numerical areas. In brief, the O-L MAT appears to be at least as effective a predictor of verbal and numerical achievement as measured by the SAT and TR as is L-T IT and MRT. Nonetheless, the reader is cautioned to bear in mind the two major restrictions on the generalizability of the study: (a) the pupils were definitely of above-average ability and very much verbally oriented; and (b) the O-L MAT and SAT are produced by the same publisher.⁵

REFERENCES

- Cooley, W. W. and Lohnes, P. R. *Multivariate procedures for the behavioral sciences*. New York: Wiley, 1962.
- Dixon, W. J. (Ed.) *BMD biomedical computer programs*. Los Angeles: University of California Press, 1967.
- Duntzman, G. H. and Bailey, J. P., Jr. A canonical correlational analysis of the Strong Vocational Interest Bank and the Minnesota Multiphasic Personality Inventory for a female college population. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 631-642.
- Groteleuschen, A. Otis-Lennon Mental Ability Test: A review. *Journal of Educational Measurement*, 1969, 6, 111-113.
- Kim, Y., Anderson, H. E., Jr., and Bashaw, W. L. Social maturity, achievement, and basic ability. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 535-543.
- Maxwell, A. E. Canonical variate analysis when the variables are dichotomous. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1961, 21, 259-271.
- Ohnmacht, F. W., and Olson, A. V. Canonical analysis of reading readiness measures and the Frostig Developmental Test of Visual Perception. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 479-484.
- Otis, A. S. and Lennon, R. T. *Otis-Lennon Mental Ability Test, Form J: Manual for Administration, Elementary I Level*. New York: Harcourt, Brace & World, 1967. (a)
- Otis, A. S., and Lennon, R. T. *Otis-Lennon Mental Ability Test, Forms J and K: Manual for Administration, Elementary II Level*. New York: Harcourt, Brace, & World, 1967. (b)

⁵ JRMCG has completed a second study on the validity of the new Otis-Lennon instrument. His recent experiment considered the topic of item bias. Interested readers should contact him directly at his new address.

THE RELATIONSHIP OF AVERAGE SCORES ON INTELLIGENCE AND READING TESTS TO PERCENTAGES OF MINORITY GROUP STUDENTS IN ELEMENTARY SCHOOLS AND HIGH SCHOOLS IN A LARGE METROPOLITAN AREA

WILLIAM B. MICHAEL, ROBERT A. SMITH, AND YOUNG B. LEE
University of Southern California

On September 30, 1969, the Los Angeles Times released for each of 435 elementary schools and 47 high schools in the Los Angeles Unified School District both the mean IQ scores and the mean grade placement scores in reading from the state mandated tests that had been administered during November 1968 as well as the percentages of minority students. Pupils in the elementary schools were given the Lorge-Thorndike Intelligence Tests, Form 1, Level D, Verbal Battery and the Stanford Achievement Test: Reading Form W, Intermediate II (level). Students in the tenth grade were administered the Lorge-Thorndike Intelligence Tests, Form 1, Level G, Verbal Battery and the Test of Academic Progress, Form 1, Reading Section.

Purpose. It was the purpose of this investigation to report for the two populations of 435 elementary schools with pupils tested in the sixth grade and 47 secondary schools with students tested in the tenth grade the degree of correlation between (a) average IQ scores and percentages of minority students, (b) average grade placement scores in reading and percentages of minority students, and (c) average IQ scores and average grade placement scores in reading. Such information might be expected to furnish a partial basis for generating a number of testable hypotheses regarding relationships between level of affluence or economic opportunity in a school community on the one hand and level of measurable scholastic aptitude or scholastic attainment on the other.

Findings. For the two populations of elementary and secondary schools, respectively, the correlation coefficients between each of the three pairings of variables were as follows: (a) IQ scores and percentages of minority enrollees, $-.826$ and $-.985$; (b) reading scores and percentages of minority members, $-.824$ and $-.890$; and (c) reading scores and IQ scores $+.962$ and $+.985$. Thus one could predict, particularly at the high school level, with a relatively high degree of accuracy the *average* scholastic aptitude scores and *average* reading performance scores in a school from knowledge of the percentage of enrollees that belong to minority groups in a given school. It should be remembered, however, that it is not uncommon to find that correlations among means of several groups are considerably higher than those found among the individual measures within the groups. Therefore, any attempt to predict individual IQ or reading scores from percentage of minority membership in a school would definitely not be warranted. The reader is left to draw his own conclusions and inferences from the data presented.

THE DEVELOPMENT OF A MEASURE OF VOCATIONAL MATURITY¹

BERT W. WESTBROOK, JOSEPH W. PARRY-HILL, JR.,
AND ROGER W. WOODBURY
North Carolina State University

VOCATIONAL maturity has come into fairly wide use as a variable presumably important in the vocational adjustment of youth (Super, Crites, Hummel, Moser, Overstreet, and Warnath, 1957; Super and Overstreet, 1960; Gribbons and Lohnes, 1968; Crites, 1965). Research to date testifies to the importance of the concept but its use is restricted by the lack of a practical, reliable, and valid instrument for measuring it. Both the Indices of Vocational Maturity (IVM) utilized in the Career Pattern Study (Super, et al., 1957; Super and Overstreet, 1960) and the Readiness for Vocational Planning (RVP) scales developed by Gribbons and Lohnes (1968) employed interview approaches which require the use of scoring manuals for assessing levels of vocational maturity. Collecting the data is time-consuming, and scoring requires a great deal of time from highly qualified personnel. The purpose of this report is to describe the development of the Vocational Maturity Scale (VMS) (Westbrook, 1970), an instrument designed to provide an objective measure of an individual's general level of vocational maturity.

A review of the literature dealing with vocational maturity suggested the following cognitive variables for which 200 multiple-choice items were constructed: (1) Related Occupations, (2) Education Required, (3) Duties, (4) Fields of Work, (5) Vocational Goal Selection,

¹This paper was supported by the research program of the Center for Occupational Education, located at North Carolina State University at Raleigh, North Carolina, in cooperation with the Division of Adult and Vocational Research, Bureau of Research, U. S. Office of Education.

(6) Vocational Problem-Solving, (7) Occupational Trends, (8) Aptitudes Required, (9) Career Alternatives, (10) Course Selection, (11) Curriculum Selection, (12) Interests, (13) Job Success Factors, (14) Vocational Planning, (15) Abilities, and (16) Values. Form A of the VMS was comprised of 100 items, five in each of the 16 areas except Education Required and Duties which contained 15 items each. Form B contained an identical number of items in each area and was intended to be equivalent to Form A. Many of the items on Form A and Form B had been administered earlier to pupils in grades 6 ($N = 1019$), 7 ($N = 2207$), and 8 ($N = 2044$) for the purpose of obtaining item analysis data which were used as a basis for the revision of items.

Form A and Form B were administered two weeks apart to a sample of 307 ninth-grade pupils in one school. Form A had a mean of 62.03, a standard deviation of 10.20, a KR-20 of .83, and a correlation of .60 with mental ability. Form B had a mean of 67.89, a standard deviation of 12.32, a KR-20 of .89, and a correlation with mental ability of .59. The correlation between Form A and Form B was .74.

To remove the unwanted factor of mental ability, each item on both forms was correlated with total scores and with scores on mental ability (Otis-Lennon). Then, the 50 items on each form having relatively high correlations with total scores and relatively low correlations with mental ability were identified. Each pupil's answer sheet was rescored on the 50 selected items on Form A and the 50 selected items on Form B.

The short version of Form A (50 items) had a mean of 36.02, a standard deviation of 5.92, a KR-20 of .77, and a correlation of .54 with mental ability. The short version of Form B (50 items) had a mean of 37.95, a standard deviation of 7.21, a KR-20 of .85, and a correlation of .49 with mental ability. The correlation between the short version of Form A and Form B (two-week interval) was found to be .65.

To obtain data on the concurrent validity of the VMS, an independent sample of 28 pupils was administered both Form A of the VMS and Gribbons' RVP, a vocational maturity instrument known to have some predictive validity. The RVP scores (average of two independent scorers) correlated .76 with scores on the VMS. The predictive validity of the VMS is currently being examined by determining whether high scorers make more appropriate vocational choices than low scorers.

REFERENCES

- Crites, J. O. Measurement of vocational maturity in adolescence: I attitude test of vocational development inventory. *Psychological Monographs*, 1965, (No. 595).
- Gribbons, W. D. and Lohnes, P. R. *Emerging careers*. New York: Teachers College, Columbia University, 1968.
- Super, D. E., Crites, J. O., Hummel, R. C., Moser, H. P., Overstreet, P. L., and Warnath, C. F. *Vocational development: A framework for research*. New York: Teachers College, Columbia University, Bureau of Publications, 1957.
- Super, D. E. and Overstreet, P. L. *The vocational maturity of ninth-grade boys*. New York: Teachers College, Columbia University, Bureau of Publications, 1960.
- Westbrook, B. W. *The Vocational Maturity Scale*. Raleigh, N. C.: Center for Occupational Education, North Carolina State University, 1970.

A PARTIAL REDEFINITION OF THE FACTORIAL STRUCTURE OF THE STUDY ATTITUDES AND METHODS SURVEY (SAMS) TEST

WILLIAM B. MICHAEL AND YOUNG B. LEE

University of Southern California

JOAN J. MICHAEL

California State College, Long Beach

ORA HOOKE

Los Angeles City College

WAYNE S. ZIMMERMAN

California State College, Los Angeles

IN a previous article three of the writers (Zimmerman, Michael, and Michael, 1970) described the results of a factorial study underlying the development of an experimental instrument Study Attitudes and Methods Survey (SAMS) Test (Michael, Michael, and Zimmerman, 1969). The purpose of this paper is to report on additional factor-analytic studies that have been carried out to refine the dimensions of the instrument.

Procedure. Both the experimental form of the SAMS consisting of 167 previously analyzed items and a supplementary inventory containing 67 new items were administered to a sample of 168 students in introductory classes at Los Angeles City College. The supplementary test form was devised to furnish a means for adding items that would yield relatively more homogeneous and reliable scales of previously identified dimensions as well as a means for separating the persistence-conformity factor into two factors of academic drive and conformity. Subsequent to a factor analysis of the intercorrelations of the 167 items and a separate factor analysis of the intercorrelations

of the 67 items in the supplementary instrument, items that exhibited low communalities and/or highly complex factorial structure were eliminated. One hundred and forty-four items of the original 167 and 43 items of the 67 new ones were subjected to a varimax factor analysis (Kaiser, 1959). Two separate solutions were obtained involving the rotation of eight and of ten principal component factors.

Results. In each of the two rotated factor analyses eight identified dimensions involving loadings above .35 on at least six item variables were described as follows: (a) *academic drive*, a form of achievement motivation involving to a large extent persevering behavior in relation to earning high marks (extrinsic motivation); (b) *conformity* (realizing teachers' expectations or meeting institutional requirements in an exacting manner); (c) *academic interest or learning affect-satisfaction* (replication of a previously found factor described as love of learning for its own sake—*intrinsic motivation*); (d) *anxiety*, or self-depreciation of one's ability to meet academic requirements including adequate performance on examinations; (e) *alienation* toward educational institutions and toward teachers and administrators as a group—a generally critical attitude toward how well the school and individuals in the power structure meet students' perceived needs and expectations; (f) *methodical and systematic approaches to study* in contrast to disorganized and slipshod techniques and poor planning; (g) *positive orientation toward teachers*—a liking of and identification with the individual teacher and with selected academic characteristics of the school environment; and (h) *manipulation* involving political shrewdness or *savoir faire* on the part of students in exerting power over instructors to gain their own ends, as if at the expense of the prestige of the teacher.

Discussion. On the basis of the findings from the factor analytic solutions, editorial revisions in certain items have been made, primarily to eliminate factorial complexity whenever possible. In particular, certain items representing a positive orientation toward the teacher were related factorially and logically to the academic interest factor and thus were merged with it. The new instrument has been refined to yield seven relatively independent dimensions consisting of 20 to 25 items in each of the following factors: (a) *academic drive*, (b) *conformity*, (c) *academic interest*, (d) *anxiety*, (e) *alienation*, (f) *method and system*, and (g) *manipulation*. New administrations of the instrument are being carried out with junior college and high

school students with the view of developing normative data and of obtaining estimates of reliabilities of each of the scales. In addition efforts are also being exerted to establish criterion-related validity for each of the scales relative to measures of academic performance.

REFERENCES

- Kaiser, H. F. Computer programs for varimax rotation in factor analysis. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1959, 19, 413-420.
- Michael, W. B., Michael, J. J., and Zimmerman, W. S. *Study Attitudes and Methods Survey*, (Experimental Form). San Diego: Educational and Industrial Testing Service, 1969.
- Zimmerman, W. S., Michael, J. J., and Michael, W. B. The factored dimensions of the study attitudes and methods survey test—experimental form. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1970, 30, 433-436.

A FACTOR ANALYSIS OF THE CPI AND EPI¹

ROBERT D. ABBOTT

California State College, Fullerton

THE California Psychological Inventory (CPI) contains 480 items and provides scores on 18 scales (Gough, 1957). Of the 18 scales, 15 were developed using the empirical approach to scale construction which consists of finding items which differentiate between a criterion group and a control group. The items in the CPI are stated in the traditional True-False first person format and some items are scored in more than one scale.

The Edwards Personality Inventory (EPI) contains 1200 items and can be scored for 53 scales (Edwards, 1966). All of the EPI scales were developed by a rational approach to scale construction. In the rational approach, emphasis is placed on constructing scales with a relatively high degree of internal consistency. The items in the EPI are stated in the third person format and the subject is asked to judge whether those individuals who know him best would answer the item True or False if they were asked to describe him. No item in the EPI is scored in more than one scale.

Both the CPI and the EPI were designed to describe personalities of "normal" individuals. Presumably normal individuals vary in the degree to which they possess various personality traits. Some of these traits may be positively correlated, others may be negatively correlated, and still others may be relatively independent of other traits. Regardless of the number of scales contained in an inventory, it is of

¹ This research was supported in part by Research Grant MH-04075 from the National Institute of Mental Health, United States Public Health Service, Allen L. Edwards, Principal Investigator.

The author wishes to thank Allen L. Edwards and Alan J. Klockars for their many helpful comments on this research.

importance to know the number of independent personality dimensions measured by the scales.

One method of determining the number of independent dimensions measured by a battery of tests or scales is to factor analyze the intercorrelations of the tests or scales.

Method. Scores on the CPI and the EPI were available for 171 female and 115 male students who participated in a test research project. Also available were scores on Edwards' (1957) Social Desirability (*SD*) scale, Welsh's (1956) *R* scale, the Marlowe-Crowne (1960) scale, and Wiggins' (1959) *Sd* scale, marker scales which have been found useful in identifying factors obtained in factor analyses of personality scales (Edwards and Walsh, 1964).

Scores on the 75 scales were intercorrelated and factor analyzed by the method of principal components. Fourteen factors with eigenvalues greater than 1 were extracted and rotated using Kaiser's Varimax. The 14 factors accounted for 71 percent of the total variance.

Results and discussion. Table 1 shows the EPI scales with the highest, positive or negative, correlation with each of the 18 CPI scales.² The table also gives the correlations of the CPI scales with the *SD* scale and the corresponding correlation of the EPI scale with the *SD* scale.

Table 2 lists the EPI, CPI, and marker scales with absolute loadings of .40 or greater on each of the 14 factors. The first two factors obtained are essentially the same as the first two factors found when the intercorrelations of the CPI scales alone are factor analyzed by the method of principal components with iterated communalities (Nichols and Schnell, 1963). Both the CPI and the EPI contain scales which measure Factor I.

With respect to Factor II, on which the *SD* scale has a loading of $-.72$, 10 of the 18 CPI scales and 2 of the 53 EPI scales have absolute loadings of .40 or greater on this factor. Factor scores on Factor II would appear to be measuring what Edwards (1970) has called the

² Tables A, B, C, and D showing the interbattery correlations of the EPI scales and CPI scales have been deposited with the National Auxiliary Publications Service. Order document NAPS 01393 from ASIS National Auxiliary Publications Service, c/o CCM Information Sciences, Inc., 22 West 34th Street, New York, N. Y. 10001. Remit in advance \$5.00 for photocopies or \$2.00 for microfiche.

TABLE 1

Highest Correlation between a CPI Scale and an EPI Scale and Their Correlations with the SD Scale

CPI Scale	r_{SD}	r_{EPI}	EPI Scale	r_{SD}
Dominance (Do)	36	76	Assumes Responsibility	27
Capacity for status (Cs)	42	-46	Shy	-41
Sociability (Sy)	46	-74	Shy	-41
Social presence (Sp)	46	-59	Shy	-41
Self-acceptance (Sa)	31	-65	Shy	-41
Well-being (Wb)	76	-46	Feels Misunderstood	-52
Responsibility (Re)	30	-33	Self-centered	-16
Socialization (So)	28	50	Conforms	05
Self-Control (Sc)	50	-50	Enjoys being center of attention	-15
Tolerance (To)	63	-55	Feels Misunderstood	-52
Good Impression (Gi)	55	59	Virtuous	28
Communality (Cm)	14	43	Cooperative	08
Ach. via Conformance (Ac)	58	53	Plans work efficiently	34
Ach. via Independence (Ai)	37	-39	Anxious about performance	-50
Intellectual efficiency (Ie)	57	-44	Feels Misunderstood	-52
Psychological Mindedness (Py)	41	-41	Perfectionist	-08
Flexibility (Fx)	02	-58	Plans and Organizes Things	10
Femininity (Fe)	-22	37	Makes friends easily	18

tendency to give socially desirable responses in self-description and what Block (1965) has called ego-resilience. In developing the EPI scales a deliberate attempt was made to minimize this personality dimension.

On Factor III, two CPI scales and eight EPI scales have absolute loadings of .40 or greater. This factor is obviously better represented in the domain of the EPI scales than in the domain of the CPI scales. With respect to Factors VII, VIII, and X, factors on which both CPI and EPI scales have relatively small loadings, there is little basis for choice between the scales in representing the factors.

There are eight factors, Factors IV, V, VI, IX, XI, XII, XIII, and XIV, on which no CPI scale has an absolute loading of .40 or greater. These dimensions of personality are, in other words, not represented by the 18 scales in the CPI.

The EPI contains 2.5 times the number of items in the CPI and requires approximately 2.5 times as long to administer. The additional testing time required with the EPI also results in approximately 2.5 times the number of dimensions of personality obtained with the CPI.

TABLE 2

CPI and EPI Scales with Absolute Loadings Greater Than .40 on the Fourteen Rotated Factors

Factor I		Factor II	
Self-acceptance	81	Tolerance	-84
Shy	-79	Well-being	-82
Dominance	79	Intellectual efficiency	78
Sociability	79	<i>SD</i>	-72
Assumes Responsibility	74	Self-control	-71
Is a Leader	74	Ach. via Conformance	-65
Articulate	73	Ach. via Independence	-62
Self presence	67	Responsibility	-56
Self-confident	64	Good Impression	-54
Capacity for Status	49	Socialization	-50
Makes Friends Easily	48	Psychological Mindedness	-50
Self-critical	-45	Feels Misunderstood	48
Enjoys Being the Center of Attention	45	Capacity for Status	-46
		Self-critical	42
Factor III		Factor IV	
Is a Hardworker	86	Avoids Arguments	78
Is a Perfectionist	81	Independent in His Opinions	-71
Persistent	77	Easily Influenced	70
Plans Work Efficiently	77	Worries About Making a Good Impression on Others	58
Plans and Organizes Things	72	Conforms	51
Motivated to Succeed	72	Critical of Others	-46
Flexibility	-54	Cooperative	43
Avoids Facing Problems	-52	Anxious About His Performance	42
Absentminded	-47	Sensitive to Criticism	41
Ach. via Conformance	41		
Factor V		Factor VI	
Kind to Others	-73	Dependent	82
Helps Others	-69	Talks About Himself	76
Careful About His Possessions	59	Wants Sympathy	61
Considerate	-57	Conceals His Feelings	-47
Feels Misunderstood	46		
Factor VII		Factor VIII	
Marlowe-Crowne	75	Flexibility	55
Virtuous	64	Psychological Mindedness	52
Good Impression	61	Communality	-51
Wiggins' <i>Sd</i>	43	Cooperative	-45
Communality	-43	Wiggins' <i>Sd</i>	-44
		Conforms	-43
Factor IX		Factor X	
Impressed by Status	77	Femininity	65
Desires Recognition	76	Conceals His Feelings	-45

TABLE 2—(Continued)

Competitive	56	Welsh R	42
		Carefree	-41
		Responsibility	41
Factor XI		Factor XII	
Has Cultural Interests	76	Feels Superior	-66
Likes to Be Alone	60	Critical of Others	-40
Intellectually Oriented	44		
Factor XIII		Factor XIV	
Interested in the Behavior of Others	-75	Active	62
Understands Himself	-66	Seeks New Experiences	51
Logical	-48		

REFERENCES

- Block, J. *The challenge of response sets*. New York: Appleton-Century-Crofts, 1965.
- Crowne, D. P. and Marlowe, D. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 1960, 24, 349-354.
- Edwards, A. L. *The social desirability variable in personality assessment and research*. New York: Holt, Rinehart, and Winston, 1957.
- Edwards, A. L. *Manual for the Edwards Personality Inventory*. Chicago: Science Research Associates, 1966.
- Edwards, A. L. *The measurement of personality traits by scales and inventories*. New York: Holt, Rinehart, and Winston, 1970.
- Edwards, A. L. and Walsh, J. Response sets in standard and experimental personality scales. *American Educational Research Journal*, 1964, 1, 52-61.
- Gough, H. G. *Manual for the California Psychological Inventory*. Palo Alto: Consulting Psychologists Press, 1957.
- Nichols, R. and Schnell, R. Factor scales for the California Psychological Inventory. *Journal of Consulting Psychology*, 1963, 27, 228-235.
- Welsh, G. S. Factor dimensions A and R. In G. S. Welsh and W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: University of Minnesota Press, 1956.
- Wiggins, J. S. Interrelationships among MMPI measures of dissimulation under standard and social desirability instructions. *Journal of Consulting Psychology*, 1959, 23, 419-427.

THE REDUCED SIZE ROD AND FRAME TEST AS A MEASURE OF PSYCHOLOGICAL DIFFERENTIATION

TED NICKEL

University of California, Los Angeles

GOUGH (1965) and Tyler (1965) both mentioned the need for a more portable, less expensive instrument with which to assess the individual's position with regard to being field dependent (FD) (Witkin, Dyk, Faterson, Goodenough, and Karp, 1962). A particularly clear description of the technical difficulties an experimenter must surmount was given by Vaught (1965) in which he described the need to blindfold subjects between trials of the RFT in order to prevent their establishing visual cues of uprightness while the experimenter was recording results of the trial.

An apparatus is described in this study which is similar to the rod and frame test (RFT) in Witkin and Asch's 1948 study. Nickel's Portable Rod and Frame Test (N-RFT), which is simple and inexpensive to construct, has criterion validity parallel to that of the more cumbersome and expensive full sized RFT.

Purpose. This experiment was conducted to determine whether the phenomena demonstrated in Witkin's large RFT could be shown through using a reduced darkened chamber and reduced rod and frame. The criterion was the Embedded Figures Test (EFT). The N-RFT scores were correlated with the EFT to see whether this reduced RFT was able to differentiate between FI and FD subjects.

Reduced RFT construction. The main box of the N-RFT was made of $\frac{3}{8}$ " plywood with the back made of two pieces of $\frac{1}{4}$ " masonite. The rotating surface on which the luminescent frame was painted consisted of a 17" and an 8" circle cut into the masonite. These

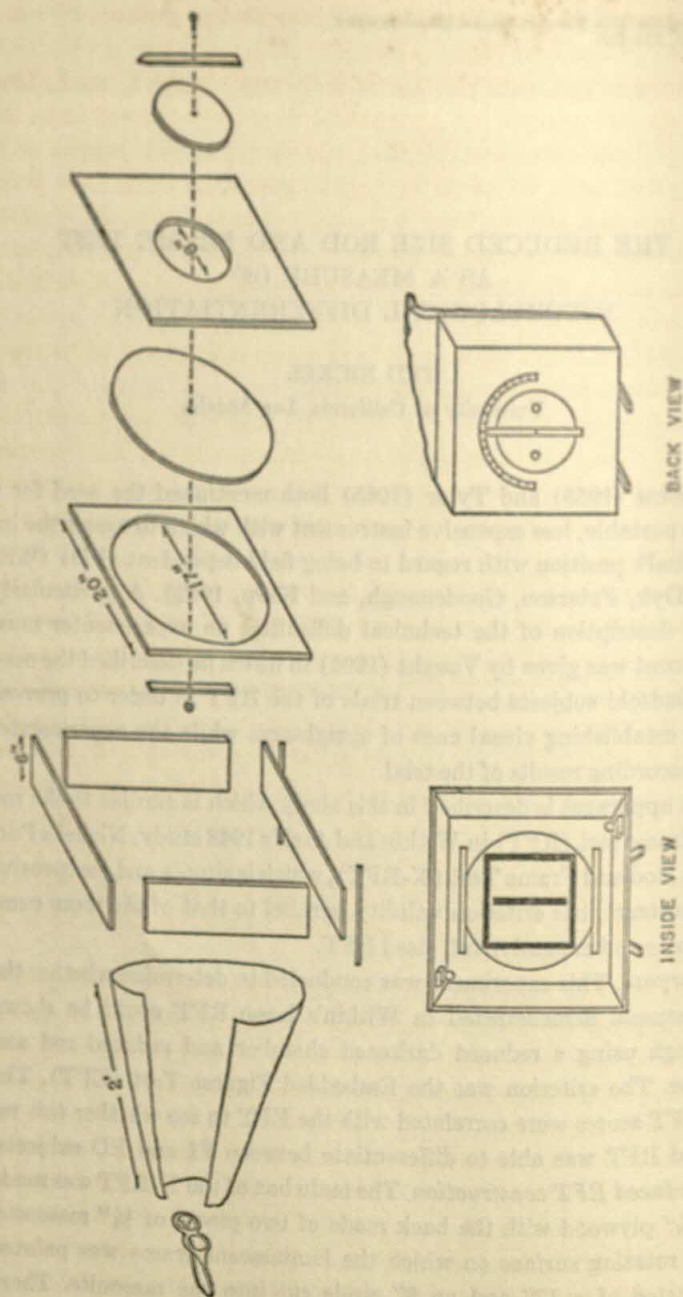


Figure 1. Construction details for Nickel's Portable Rod and Frame Test

two circles were glued together and then held in position by wood blocks.

For the hood, a black plastic material (.006 in. thick), usually used to cover construction and agricultural goods was tailored to fit the box. The mask was from a set of U.S. Army surplus goggles ordinarily used with red lenses for night vision adaptation. The lenses were removed, and the hood was attached to the soft rubber frame of the goggles. Electrician's plastic tape was used to join the seams and to attach the plastic hood to the box as well as to attach the hood and goggles.

Two 15 watt light bulbs were fixed in opposite corners to "charge" the luminescent surfaces between trials.

The luminescent rod was 1" \times 8", the frame 12" square with the strip 1" wide. The frame was painted directly onto the large rotating surface. (The luminescent paint used was made by Lawter Chemicals Incorporated of Chicago, Illinois.) The rod rotated independently on its own shaft.

Method. The N-RFT was administered following Witkin's (1949) instructions. However, only one set of eight trials, rather than the original three sets of eight trials each, was given. According to Witkin et al. (1963), these extra two sets, which would have involved tipping the subject 28° left and right, would not measurably affect validity. The N-RFT deviation score for each *S* consisted of the sum (over eight trials) of the absolute deviations in degrees of his placement of the rod from the true vertical. Following the N-RFT, subjects were given the EFT by using Jackson's (1956) revision of the EFT procedure. Jackson's revisions have shortened the testing time by half and still have maintained more than .96 reliability.

The subjects ranged in age from 10 to 19 years. Of the 38 girls, 8 were in grade school, 10 in high school, and 20 were college freshmen. Of the 42 boys, eight were in grade school, seven in high school, and 27 were college freshmen. The college subjects received course credit for their participation.

Results. A test for reliability (internal consistency, odd-even, Spearman Brown corrected) yielded $r = .92$. Concurrent validity coefficients of the N-RFT with the EFT as criterion produced $r = .70$ for all 80 *Ss* ($r = .74$ for the 38 girls, and $r = .50$ for the 42 boys). All three coefficients were highly significant ($p < .001$).

A *t*-test of boys versus girls performance on the RFT resulted in $t = 3.25$ ($df = 78$, $p < .01$, two-tailed). The girls' mean deviation score of rod placement from true vertical was 40.58 degrees arc, while the boys mean deviation score was 19.07.

Discussion. The reliability (internal consistency) of the N-RFT compares favorably to that obtained with the full size version (Witkin, 1962). The validity of the N-RFT for measuring FI-FD was clearly established. While subjects' mean deviation score on the N-RFT was about half that found by Witkin (1949), the N-RFT clearly maintains the ability to discriminate FI from FD subjects.

Some of the subjects were markedly affected during the N-RFT session. Typical comments were: "My stomach began to feel upset," and, "When the frame was shifted, it was like going over a hump in the road very fast in a car," or "I had to close my eyes once in a while to keep from getting upset!" It was felt that the use of a flexible plastic hood prevented the subject from obtaining positional cues, as would be the case if the subject's head was immobilized. The absence of positional cues may have led to the marked effect of the N-RFT on the subjects as well as to the high agreement between the EFT and N-RFT scores. The tendency toward nausea under conflicting conditions (visual versus gravitational) was reported by Witkin (1949). The significant difference between boys and girls in rod placement, with girls having a higher deviation score, also agrees with findings of Witkin et al. (1962).

Bercovici (1970) has used the N-RFT to assess 102 subjects' relative position in terms of psychological differentiation. The N-RFT was successful in defining two personality subgroups, FD and FI. The FI group was found to be significantly more aggressive.

It is felt that the results of this validation study indicates that it is possible to use the rod and frame approach for measuring psychological differentiation without the burdensome technical problems presented by the full size instrument. The gain in convenience, it was shown, can be obtained with little or no loss of information through the use of Nickel's Portable Rod and Frame Test.

Summary. Witkin's Embedded Figures Test (EFT) and Nickel's Portable Rod and Frame Test (N-RFT), a version of Witkin's Rod and Frame Test (RFT) were presented to 80 elementary through college age subjects. The N-RFT is completely portable, the size of a suitcase, and does not require a darkened room. Concurrent validity

with the EFT as the criterion measure, yielded $r = .70$. A t -test indicated a significant sex difference beyond the .01 level as would be predicted from Witkin's findings. It was felt that the N-RFT could be used in addition to the EFT for research with psychological differentiation.

REFERENCES

- Bercovici, A. M. Children's aggression, television viewing, and psychological differentiation. Unpublished master's thesis. University of California, Los Angeles, 1970.
- Gough, H. A. Embedded Figures Test In O. K. Buros (Ed.), *The Sixth Mental Measurements Yearbook*. Highland Park, N. J.: The Gryphon Press, 1965.
- Jackson, D. N. A short form of Witkin's embedded-figures test. *Journal of Abnormal and Social Psychology*. 1956, 53, 254-255.
- Tyler, L. Embedded Figures Test In O. K. Buros (Ed.), *The Sixth Mental Measurements Yearbook*. Highland Park, N. J.: The Gryphon Press, 1965.
- Vaught, G. M. The relationship of role identification and ego strength to sex differences in the rod-and-frame test. *Journal of Personality*, 1965, 33, 271-283.
- Witkin, H. A. The nature and importance of individual differences in perception. *Journal of Personality*, 1949, 18, 145-170.
- Witkin, H. A. and Asch, S. E. Studies in space orientation: IV. Further experiments on perception of the upright with displaced visual fields. *Journal of Experimental Psychology*. 1948, 38, 762-782.
- Witkin, H. A., Dyk, R. B., Faterson, H. F., Goodenough, D. R., and Karp, S. A. *Psychological differentiation: Studies of development*. New York; Wiley, 1962.

BOOK REVIEWS

MAX D. ENGELHART, Editor
Duke University

HENRY MOUGHAMIAN, Assistant Editor
City Colleges of Chicago

<i>Andrew and Moir's Information-Decision Systems in Education.</i> ROBERT A. SMITH	563
<i>Brown's Appraisal Procedures in the Secondary Schools.</i> CARL A. CLARK	565
<i>Brown's Measurement and Evaluation.</i> MAX D. ENGELHART ..	569
<i>Downie and Heath's Basic Statistical Methods.</i> WILLIAM M. STALLINGS	570
<i>Downie's Study Guidebook to Accompany Basic Statistical Methods.</i> WILLIAM M. STALLINGS	570
<i>Guertin and Bailey's Introduction to Modern Factor Analysis.</i> W. TODD ROGERS AND LARRY R. NELSON	572
<i>Kelly et al. Research Design in the Behavioral Sciences.</i> JOHN L. WASIK	576
<i>Lyman's Test Scores and What They Mean.</i> J. STANLEY AHMANN	579
<i>Stern's People in Context.</i> HOWARD G. MILLER	581

Gary M. Andrew and Ronald E. Moir. *Information-Decision Systems in Education*. Itasca, Illinois: F. E. Peacock, 1970. Pp. xii + 177. \$5.75 and \$3.95 (paperback)

The stated purpose of this book "... is to present the student and practitioner of education administration with an integration of (a) a formal structure for decision making and (b) the methodology for designing an information system to support the decision making function" (p. vii). While these objectives are seemingly foreign to educational and psychological measurement, a close examination of the decision-making function indicates that a key component consists of a set of objectives which "... should be explicit enough to enable one to measure whether the objective is being realized" (p. 8). In this context the close relationship between systems analysis and summative evaluation seems obvious. Both are concerned with specifying educational objectives, determining alternative methods of achieving the objectives, and evaluating which of the alternatives are most efficient in achieving the specified objectives.

Since the book is intended to present an overview of a "... continuum from information to decision that should be recognized when managing an educational system" (p. vii) it would seem reasonable to assume the book would develop such a continuum. Unfortunately this is not the case. The initial discussion consists of a brief expose of problem solving and decision making. One of the items in this discussion is a consideration of objectives which includes the following: "In education it may be argued that this is not true since everyone knows that the objective of education is 'to provide educated citizens'" (p. 8). This example is then used to illustrate the need for specifying a set of criteria to evaluate objectives. It is suggested that the approach used by Bloom et al. in the *Taxonomies* would develop more precisely stated objectives than the procedure detailed in this treatise.

The applicability of systems analysis to education is next developed. The major rationale for this applicability is the process of the aerospace industry in solving "... a great number and variety of unprecedented system design and development problems associated with national defense" (p. 21). Quite apart from the undenied successes (and possibly some failures as witness the continuing eight year TFX-F111 airplane controversy) there is the basic assumption that national defense and education are sufficiently similar to be correctly subjected to the same sort of analysis. The authors argue that systems analy-

sis needs an objective which is something or things that can be verified as being accomplished by the system (p. 22), "what is really required is an analysis of the *whole* educational structure *before the fact*" (p. 25), and finally "the problem (of defining precise educational objectives) involves differing value judgments among both educators and the community on what the most important goals and purposes of education are . . ." (p. 29). In the context of a pluralistic society these three statements are, to this reviewer, seemingly incompatible. Apparently the authors recognize this incompatibility when they state "... we need to know the relationship between what is *done* in school and what students *learn* in school. The systems approach to education helps to focus clearly on the unknown relationships and gives a clear indication of the directions educational research should take in the future, and what answers we should be seeking" (p. 29). This abrupt degeneration of systems analysis as the means of solving the problems of education to a procedure for developing guidelines for educational research is, to say the least, surprising (given the promise of funds to undertake the research, the leaders in the field of education could undoubtedly provide the suggested directions for research and at the same time shed a great deal of light on the authors' "unknown relationship").

Model concepts are next introduced and a sample flow chart (p. 33) is presented. It is doubtful that the subject of the flow chart would be of interest to the intended audience. Further, without some detailed explanation of flow charting techniques the advantages of the procedure would not be readily apparent to the uninitiated. At the same time this reviewer would question the necessity of including an appendix of flow chart symbols which would only be of use to individuals responsible for developing computer programs.

Fortunately models are briefly considered; attention is next directed to variables. "A variable in a system is a well-defined attribute which describes the value or condition of certain aspects of the system. It is important in systems analysis and model building that the pertinent variables in a system be described and understood. The determination of what is pertinent is a highly subjective art form because there are degrees of effect that various variables have in a system. The most important thing to remember about variables is that they must be defined in such a manner that they are completely unambiguous and have the same meaning to all individuals concerned with the system" (p. 35). One could only wish that the concept "variable" had been subjected to the last quoted requirement.

One particularly difficult problem, educational objectives, is next considered and disposed with "... we take the position that achievement in basic subjects is the most widely accepted and the most important dimension of educational output" (p. 43) and "... systems performance can be evaluated by the aggregation of the individual

scores (or batteries of standardized achievement tests)" (p. 45). If this position were accepted, which is, incidentally, contrary to the authors' position on page 29 (quoted above), then education could quickly, with or without systems analysis, put its house in order. This reviewer would argue that neither education nor the community which it serves would be willing to accept such a narrow definition for the output of the educational system.

The remaining chapter details in a straight forward manner procedures for implementing an information system which details the rationale for one of the two case studies, "An Integrated Educational Information System" (p. 132ff). This discussion and the case study would be most valuable to any individual interested in the development of a comprehensive information collection and dispensation system. One could only wish the consistency which is so admirably evident in these sections concerned with the information systems had been applied to the discussions of systems analysis and decision theory.

ROBERT A. SMITH

University of Southern California

Donald J. Brown, *Appraisal Procedures in the Secondary Schools*. New Jersey: Prentice Hall, Inc., 1970. Pp. iii + 182. \$5.95, \$2.95 (paperback).

"Appraisal Procedures in the Secondary Schools" is a title that suggests either a survey of appraisal procedures presently used in high schools or an instructional book on how to appraise high school students. Apparently the latter is the principle intention since the author, in the preface, says, "The book is intended to help teachers, or prospective teachers, acquire and understand the principles and procedures that will allow them to do a more effective job in evaluating student achievements." However he also states that much of the content and structure of the book was provided by actual teacher and student comments from over 400 taped, structured interviews. Since there is little indication of how many persons expressed the viewpoints given in the comment, except for occasional use of "typical," "many," or "one teacher," it is really not a survey of current procedures.

The use of the comments from the 400 taped interviews can be considered as illustrative, and perhaps as a readability technique. The latter possibility is suggested by the putting of appraisal or measurement techniques in a form easy to swallow, with very little mathematical and what the author calls, "psychometric jargon." The small size of the book, some of the chapter headings, the style, and a good deal of the material suggest an attempt to sugar-coat what students presumably take to be the bitter pill of educational measurement.

Following an introductory chapter on the organization and the general methods to be used, the author proceeds with a chapter called, "The Better I Teach, the Better My Students Will Do on My Examinations." The author then gives some methods, not for better teaching, but for helping students do better on examinations. These methods include preparing students for examinations, setting up course objectives that can be tested, giving suggestions to students that will ease emotional tension, testing students frequently, and carefully going over examinations that have been taken. One should teach so that students do better on examinations, and students should be helped in every way possible to do better on them. Under such conditions teaching could become almost an examination coaching process. Better that examinations should be eliminated altogether than that this should come to pass.

The best of what a student can get from a class probably cannot be measured by any test, less likely a teacher-made test. All we can hope for is that a test score is to some extent correlated with the important factors that cannot be measured. These factors include such things as, interest of the student in the course and in what it means to his future; the fellowship of teachers and students working together; skills in locating and using resource persons and materials; and use of knowledge and skills from the course in conversation and discussion in and outside of the class. Such things could be deviated or eliminated through concentration of teaching on the improvement of scores on the very imperfect classroom measuring instruments that are likely to be used.

The third chapter is called, "How do you Know Whether an Examination is Good"? The question is not answered. For a test to be good, it is pointed out, it must be reliable and valid, and also a brief discussion of reliability and validity is undertaken on a very elementary and non-mathematical level, with some suggestions for improving reliability and validity.

The chapter on "What Should Classroom Test Measure"? gives a fairly specific and helpful way of constructing a classroom test. Construction begins with working out the specific objectives of what is to be taught and evaluated. From these is made an outline of the content of the teaching unit, with the amount of time to be spent in each area. The examination questions are then constructed in these areas, with the proportionate number of questions in each area approximating the proportion of time spent in each area. So far so good, but in the detailed example given, the test content, as shown in the "test plan," is taken directly from the unit outline.

The specific facts, the terminology, the ideas, all are presented in the outline, which is for a two weeks instructional period, at the end of which the examination is to be given. In two weeks of instruction with fifty-one minutes per day surely more than two pages of ma-

terial will be covered. A student having this outline, and all students should have it, could memorize everything to be on the test in fairly short order. A good cramming session should easily take care of that test.

If an outline is to guide learning, then the learning the student has been guided into is what is of primary importance, not the guide itself. The general facts and statements of the outline are a guide to the more specific statements, demonstrations, and explanations of basic course content. Deduction as well as induction is an important part of learning, and the student should be able both to develop a generality and to support a generality with specific factors and details. Consequently, in testing it is important to discover whether or not a student has these specifics with which he can demonstrate and support the general statement. It is for this reason that all material in a course is important, if it belongs in the course at all, and the outline is only a very small bit of the course content.

The two following chapters are concerned with the construction of essay and objective type tests. Some good points are given, though not everyone would agree that the essay examination is a more valid measure of achievement than the objective type, or that one should be sure that one is measuring things other than knowledge or specific facts with an essay test, since these are best measured with objective tests. The importance of proof-reading essay question answers is affirmed, but not for objective examinations, for which it is at least as important. It often happens on a true-false or multiple-choice item that the student inadvertently marks a choice he had not intended or misreads a question. As a matter of fact the author himself makes such an error in one of his examples, on item construction:

(T) F 9. Benjamin Franklin lived a long time ago.

Better:

(T) F 9. Benjamin Franklin lived in the sixteenth century.

Even so, there is much that should be helpful to students in these chapters, as well as in the following chapter on constructing and analyzing an objective test. In the discussion on item analysis, however, one might question the use of the term "validity index" for the difference in proportion of right answers between the upper and lower groups, which ordinarily is referred to as a kind of item discrimination index. In this chapter he states again, though somewhat differently the purpose of testing:

Since the major purpose of a classroom exam is to discriminate among how much each student learned in a unit of instruction, it is desirable to have a test that yields a wide range of scores.

Translated, this statement recognizes that a classroom test seeks to differentiate students as to their relative amounts of learning, but does not measure the actual quantity of learning achieved by each

student. His statement, therefore, is somewhat at odds with the earlier one that the purpose of the classroom examination is to determine how much the students have learned, which implies a scale of a higher order than those used in the classroom.

The last quoted statement also does not agree with his earlier assertion that how well a student does on a test is an indication of how well the teacher has taught. If there is a wide dispersion of scores, then, if the latter is true, the teacher has taught some students well and some poorly. In other words the teacher who has constructed a good test is revealed as a good teacher for some students and a poor teacher for others. If a teacher really wants to know how good or bad he is, he better let someone else do the evaluating.

Then follows a little chapter on statistics, which is entitled: "The Mystery Hour: A Few Statistics." It is inexcusable to present statistics to students in this way. Too often statistics seem to them an abracadabra pronounced over numbers to make them good or bad. To foster such a superstition through the use of this chapter title, in place of attempting to bring about an acceptance and understanding of the use and value of statistics is to defeat the cause of improved measurement. The few statistics that are given are presented so inadequately that an hour spent on this chapter would truly be a mystery hour.

In the chapter on assigning grades there is some good discussion, but the measurement student who turns to it in the hope of receiving much help in assigning grades is likely to be disappointed. Perhaps he need not be too much concerned, however, for the author states:

Although there is some evidence to support it, this chapter's intent is not to condemn all grading practices and demand their immediate revision, but to indicate that marking procedures are necessarily only as good or as bad as the teacher who is trying to apply them.

This statement implies that there are no good or bad ways of grading, there are only good or bad teachers. If you are a good teacher, use any method of grading you want, and if you are a bad teacher you might as well use any method too, since then in any case it will be bad.

This chapter discusses grading by inspection, using gaps to indicate grade separation, and grading on a curve. The weaknesses of these methods are pointed out, and the author seems to prefer the "modified curve," in which other factors than the test score is taken into account, such as homework and class participation, though he fails to indicate how each of these is to be evaluated separately. The procedure he suggests for computing a final grade consists of adding the weighted grades of examinations, papers, reports, and class recitations: but again, no analysis is made of how each separate grade

might be assigned, except for the methods of test grading of which he is highly critical.

In the tenth chapter, "Published Tests: An Evil or a Blessing" there is a good discussion of intelligence testing on a very general and elementary level, giving some important precautions for interpretation of IQ's. There is no specific help, however, in interpreting obtained IQ's. No mention is made of the terms "gifted," and "mentally retarded." It is emphasized that the interpretation of scores should always take into account measurement error, but there is no mention of how it should be taken into account. There is also a brief discussion of standardized achievement tests, of student motivation, ranking such tests, and of telling parents and students the results.

In the final chapter, "A Look into the Future," instructional television, programmed instruction, computer assisted instruction, and tele-instruction, and tele-lecture are discussed very briefly. The book concludes with what is apparently the theme of the book, "The need for better evaluation of student achievement, now as in the future, is the teacher's." One might add to this that one key to better teaching is better evaluation of student achievement.

CARL A. CLARK
Chicago State College

Frederick G. Brown. *Measurement and Evaluation*. Itasca, F. E. Peacock, 1971. Pp. xiv + 198. \$3.95 (paperback).

According to its author, this book "was written primarily for an introductory course in educational psychology" and to emphasize "aspects of measurement and evaluation most pertinent to the classroom teacher." It should be evaluated with these purposes in mind.

The introductory chapters are concerned with the functions of tests and the basic qualities of measuring instruments—standardization, objectives, content sampled, directions to insure uniform testing conditions, scoring, consistency or reliability, and validity. Both reliability and validity concepts are explained and illustrated quite adequately considering the level of its intended student audience. The paradigms contrast graphically test-retest reliability, or the coefficient of stability, equivalent forms reliability, the coefficient of equivalence, and the coefficient of stability and equivalence. Similar criterion-related validity is explained and illustrated graphically. These paradigms would seem excellent teaching devices. Internal consistency and split-half coefficients, construct validity, and content validity are briefly and reasonably adequately explained. The reviewer does not agree that construct validity should be of "little concern to most classroom teachers" although the author is probably justified in saying that it is of limited concern.

The discussion of the interpretations of test scores in terms of

norm data and types of derived scores is well explained and illustrated. Commendably critical mention is made of criterion-referenced measurement. (This reviewer deplors the term criterion-referenced tests because of probable confusion with criterion-related validity. The idea of such interpretation of test data is almost as old as the measurement movement although tests have too seldom been designed to promote its accomplishment.)

The chapter on classroom tests contains excellent "GUIDELINES FOR CONSTRUCTING MULTIPLE-CHOICE ITEMS" and "GUIDELINES FOR CONSTRUCTING TRUE-FALSE ITEMS" though this reviewer does not believe that teachers should be encouraged to construct the latter. The discussion of matching items is most inadequate. There could be much more explanation and illustration of matching exercises suitable for use with hand-scored, or machine scored, answer sheets. There should be some discussion of keylist exercises. These are especially useful with reference to quoted material in measurement of intellectual skills.

The "GUIDELINES FOR CONSTRUCTING ESSAY QUESTIONS" and the "COMPARISON OF ESSAY AND OBJECTIVE TESTS" are excellent "EXHIBITS." They are accompanied by satisfactory discussion of advantages and limitations.

The chapter on analyzing test scores and test items, the explanations of the computation of percentile ranks and standard scores, and introduction to elementary descriptive statistics in an appendix are unusually clear and complete for so small a book.

The later chapters of the text deal successively with examples of well-known standardized achievement, general scholastic aptitude, and vocational aptitude batteries. The kinds of exercises in each of the subtests of each of the batteries are illustrated by ample quotations of their practice exercises.

The final chapters of the book contain discussions of grading, the evaluation of instruction, and include a summary or "recapitulation."

While the book has some minor limitations, it should serve the purpose for which it was written. It is the opinion of this reviewer that it might also serve, when accompanied by one of the recent books of readings in the field, a good choice for an undergraduate course in educational measurement.

MAX D. ENGELHART
Duke University

- N. M. Downie and R. W. Heath. *Basic Statistical Methods*. (3rd Edition). New York: Harper and Row, 1970. Pp. xi + 356. \$9.95.
N. M. Downie. *Study Guidebook to Accompany Basic Statistical Methods*. (3rd Edition). New York: Harper and Row, 1970. Pp. 125. \$2.95 (paperback).

Several years ago when this reviewer began teaching introductory

statistics, a book often recommended was *Basic Statistical Methods* by Downie and Heath. Obviously, any text which has now gone through three editions must be satisfying the needs of many instructors and students. *Basic Statistical Methods* (3rd ed.) does have its several virtues. The book is well written, easy to read, and makes almost no demands upon the mathematical abilities of the student. One might also classify this text as a "cookbook," a rubric not intentionally opprobrious but one which does indicate that many topics are not explored in depth. Problems (and their answers) are provided for each chapter.

Basic Statistical Methods is composed of 18 chapters which are in turn divided into three parts. "The first nine chapters present descriptive statistics, and the next seven consist of an introduction to statistical inference. The third part consists of two unrelated chapters, first an introduction to test theory and construction, and second, a look at the more frequently used distribution-free statistical tests." (xi)

Technically, there is little to quibble with in the first part. Some topics are discussed which, to this reviewer's mind, might well have been omitted. For example, the computation of square roots and of grouped data techniques (even to the extent of giving Charlier's check and Sheppard's correction) seem out of place in a modern statistics book. And purists may be bothered by the use of N rather than $N - 1$ in the product-moment correlation formula. Nevertheless, this section is readable and competent. Indeed, the first nine chapters (encompassing an introduction, a mathematical refresher, frequency distributions, averages, variability, standard scores, product-moment correlation, other correlational techniques, and linear regression) are well suited to accompany an introductory course in measurement.

The second part is the weakest section of the text. Weaknesses include anachronistic advice, inappropriate examples, and errors of interpretation. The probability chapter is traditional, i.e., similar to high school texts circa 1950. A welcome inclusion is the chapter on sampling. Unfortunately, this otherwise good chapter is marred by its treatment of confidence intervals. With a set of sample data, confidence is given a direct probabilistic interpretation (p. 164). The chapter on hypothesis testing offers a non-standard convention relative to significance: when the authors discuss two tailed tests, they speak of *level* of significance; when they present one tailed tests, they refer to *point* of significance. In choosing between t and Z test statistics in an independent groups situation, the authors advocate using the criteria of sample size and homogeneity of variance. No explicit consideration is given to whether the population standard deviation is known or estimated, although the formulas presented (t and Z) in effect contain variance estimators. There are no discernable notational distinctions between sample variances and variance estima-

which Kaiser (1963) called the most psychometrically defensible approach extant, is an egregious omission.

The chapter devoted to the number of factors problem also lacks a modern orientation. The reader is not informed of contemporary methods of determining the number of factors. The statistical tests applicable in maximum likelihood solutions are not mentioned. This perhaps may be accounted for by the general absence of any discussion of maximum likelihood methods. However, the omission of the psychometric bounds recommended by Guttman (1954) and Kaiser (1960, 1963) is inexcusable. The authors' recommendation not to rotate principal components is intolerable in light of the successful application of what has become known as the "Little Jiffy" method of factor analysis.

The Harris-Kaiser (1964) method of oblique transformation is not mentioned. The treatment of factor scores is incomplete and outdated, with more recent writings and recommendations completely ignored (e.g., Harris, 1967).

Granted, then, that the text is not an introduction to modern factor analysis, might it assist the reader in understanding the basic process of factor analysis? The answer is a qualified, No. The authors' emphasis on geometrical representations is helpful in understanding the concepts of centroid, factor loading, and orthogonal and oblique (reference and primary) factors. Thurstone's cylinder and box problems are handled nicely and used in one enlightening chapter dealing with the meaning of factors (Chapter 5). In the preface, the authors wrote: "We have written this book for the mathematically unsophisticated reader because he has the most difficulty in finding out about factor analysis. . . . We have tried to restrict the topics and depth of coverage to the needs of the neophyte reader" (p. viii). However, the reader will find himself thrust immediately into short, confusing chapters. The introductory chapters depend on some prior knowledge of factor analysis which is particularly distressing since the text is intended to be an introduction to factor analysis. Basic concepts such as method and scale dependencies are not presented. The pioneering works of Spearman, Thurstone, and Holzinger are not discussed. This is a major weakness, especially for the type of reader discussed in the preface. The classic works provide the beginning student with a substantive base of the original theory and purpose of factor analysis, giving the basic associations necessary for the tyro.

Terms certain to be unfamiliar to the novice are used throughout the text without appropriate introduction (e.g., simple structure (p. 15), factor matrix (p. 15), reference factors (p. 42), fallible data (p. 43), gross rank (p. 26), determinant (p. 26), Gramian (p. 153), Maxplane (p. 106)). The reader is advised of this in the preface and offered numerous apologies; only a high tolerance for

ambiguity, tedium and resultant frustration will enable the reader to continue with his effort. The frustrated reader can find solace with the typist of the text who at the end of Chapter 2 "found the next chapter particularly tedious but was delighted with the meaningfulness of Chapter 5" (p. 35). One wonders what the typist's opinion was of Chapter 4. Guertin and Bailey, in referring to structure and pattern values in connection with oblique solutions, add to the confusion already present by adopting their own set of labels (p. 106), and using such terms as factor loading, factor matrix, and factor in the broadest sense possible (p. 104). On another occasion, the authors incorrectly refer to the correlation between a continuous dependent variable and a dichotomous independent variable as a biserial r rather than as the correct point-biserial r (pp. 199-200).

Guertin and Bailey's informal, casual writing style ("You say, 'Hooray! I'm sure glad it didn't come out to .90 because it should be measuring something different'" p. 225) may appeal to some readers, but insult others. Some direct quotes are made without adequate references to the author or page number (e.g., p. 222, p. 202). Several comparisons are made between various derived solutions using a "factor loading code." The code is incorrectly explained on p. 131; in several tables coded information is not clearly represented (e.g., p. 150).

By their own admission, Guertin and Bailey have not provided a scholarly reference on factor analysis. One reads on the dust jacket that "the book was developed as a text for the senior author's classes but it parallels much of what has been explained to colleagues in consultations about how to analyze their data." Perhaps this book is appropriate for the aforementioned colleagues. However, novice mathematical thinkers could do no better than heed the suggestion of the authors to begin their study of factor analysis from a more mathematical text such as Harman (1967). The nonmathematical beginner might more profitably begin with Thurstone (1947), or Cattell (1952, p. 1-108), then turn to Harman (1967). To understand modern methods of factor analysis, the reader must turn to references such as those listed below. A good, *introduction to modern factor analysis* has yet to appear.

REFERENCES

- Cattell, R. B. *Factor analysis*. New York: Harper, 1952.
 Cattell, R. B. *Handbook of multivariate experimental psychology*. Chicago: Rand McNally, 1966.
 Guttman, L. Image theory for the structure of quantitative variates. *Psychometrika*, 1954, 18, 227-296.
 Guttman, L. Some necessary conditions for common-factor analysis. *Psychometrika*, 1954, 19, 149-161.

- Harman, H. H. *Modern factor analysis* (2nd ed.). Chicago: University of Chicago Press, 1967.
- Harris, C. W. Some Rao-Guttman-relationships. *Psychometrika*, 1962, 27, 247-263.
- Harris, C. W. Some recent developments in factor analysis. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1964, 24, 193-206.
- Harris, C. W. On factors and factor scores. *Psychometrika*, 1967, 32, 363-379.
- Harris, C. W. and Kaiser, H. F. Oblique factor analytic solutions by orthogonal transformations. *Psychometrika*, 1964, 29, 347-362.
- Jöreskog, K. G. Some contributions to maximum likelihood factor analysis. *Psychometrika*, 1967, 32, 443-482.
- Kaiser, H. F. Psychometric approaches to factor analysis. *Proceedings, ETS Invitational Conference on Testing*, 1964. Princeton, N. J.: Educational Testing Service, 1965.
- Kaiser, H. F. The application of electronic computers to factor analysis. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1960, 20, 141-151.
- Kaiser, H. F. Image analysis. In C. W. Harris (Ed.), *Problems in Measuring Change*. Madison, Wisc.: University of Wisconsin Press, 1963. Chapter 9, pp. 156-66.
- Kaiser, H. F. and Caffrey, J. Alpha factor analysis. *Psychometrika*, 1965, 30, 1-14.
- Rao, C. R. Estimation and tests of significance in factor analysis. *Psychometrika*, 1955, 20, 93-111.
- Thurstone, L. L. *Multiple-factor Analysis*. Chicago: University of Chicago Press, 1947.

W. TODD ROGERS

LARRY R. NELSON

Laboratory of Educational Research
University of Colorado

- Francis J. Kelly and others. *Research Design in the Behavioral Sciences; Multiple Regression Approach*. Carbondale, Ill.: Southern Illinois Press, 1969, pp. xiii + 353. \$6.00.

The title of this text would lead perspective readers to believe the contents of the text will provide the coverage of both design and analysis procedures. While the text is concerned with the use of multiple regression procedures for data analysis purposes, there is very little in the way of design principles and procedures. Indeed, there appears to be implicit in the text the feeling that design considerations are not important as long as the researcher has available a multiple regression program and a computer to run it on.

As introduced by the authors, the multiple regression approach is perceived as a means of allowing the researcher to develop a

statistical model which will reflect the research question originally asked. The text begins with the presentation of a structural model to account for human behavior according to a three level classification scheme of independent or predictor variables: (1) within person; (2) focal stimuli characteristics, and (3) context characteristics (i.e., environment). It is then suggested that these variables can be arranged in a multiple regression equation to represent the hypothesized functional relationship in much the same manner economists utilize least squares procedures to specify theories. Chapter II provides a review of inferential statistics with a focus on the concepts of variance and error sums of squares. Chapter III presents a brief discussion of the representation of variables within a vector framework. It is of interest to note that sets of vectors are utilized to describe the design of interest while most treatments of least squares analysis use a matrix to specify the experimental design of interest. This chapter also presents, in a non-mathematical fashion, some basic concepts of vector algebra. The fourth chapter compares the regression analysis and the analysis of variance algorithms for testing stated research hypotheses.

In the fifth chapter, the reader is introduced to use of the computer program "LINEAR" for obtaining answers to the questions posed by the researcher. This program is derived from the PERSUB computer subroutine system developed by Joe Ward, Jr. at the Personnel Research Laboratory, Lackland Air Force Base, Texas. Since Joe Ward and Bob Bottenberg were responsible for the development and dissemination of this particular approach to data analysis as well as the program reproduced in the text, it would have been appropriate to at least acknowledge the lineage of the program. It should also be noted that the same approach to data analysis can be followed with any multiple regression computer program. The major attraction of LINEAR are the provisions for data manipulation and for the tests of hypotheses via comparison of full and reduced models.

Roughly one third of the text is contained in Chapter VI where the reader is quickly introduced to the use of regression analysis to solve research questions of a fairly complex nature. Methods are presented to test effects of: (1) independent variables in a categorical and/or continuous forms; (2) interaction and non-linear forms of curvilinearity; and (3) higher degree polynomials. Several data topics in Chapter VI are treated in both intuitive and formal presentations. The intuitive approach indicates what is the outcome of arbitrarily selecting partial regression weights while the formal presentation gives the mathematical basis for solving the regression equation developed to answer the specified hypotheses.

Chapter VII presents an extension of the regression analysis approach to problems where statistical adjustment through the use

of co-variables is required. Also included is a section on the relationship of multiple regression analysis to other multivariate procedures such as discriminate analysis, factor analysis and canonical correlational analysis. Somewhat surprising is the lack of discussion concerning the relationship between multiple regression procedures and the more general multivariate analysis of variance approach to data analysis.

The last chapter (VIII) entitled "Special Considerations Regarding Multiple Linear Regression Analysis" presents a potpourri of recommendations and comments concerning the use of multiple regression programs such as LINEAR. Included are topics related to handling missing data, accuracy of iterative solutions, simulating an analysis of variance via regression analysis, repeated measures, and data transformation. Also included, as an afterthought it appears, are algebraic proofs relating equivalence of different tests of significance and R^2 and variance of predicted scores.

According to the authors, the text was written for prospective and practicing researchers in the behavioral sciences. Since it is likely that prospective readers may have differing levels of skills in the areas covered in the text, the extensive use of descriptive chapter subheadings would provide a person with an opportunity to pick and choose among topics to be studied. The availability of problem sets with answers would tend to suggest this book could be used in a self study format. Somewhat surprising is the lack of statistical tables. While LINEAR does provide a calculation of the approximate probability for the obtained F statistic, it is likely that some prospective users of this text may choose not to implement LINEAR, particularly in view of the accuracy level provided by the iterative process of the program (see pp. 25-254). By using a local regression analysis computer program, a researcher can easily calculate an F ratio by hand and refer to an F distribution table to determine the significance of the obtained F value.

This reviewer has used the approach of Kelly et al., for introducing students in a first year experimental design course to the concept of general linear models. Since most of these students have been exposed to classical research design procedures prior to this course, they have no trouble seeing the isomorphism between the analysis of variance and regression approaches for simple one way and factorical designs with two levels per factor. However, the students appear to have a great deal of trouble in seeing the equivalence of the two procedures when the attempt is made to extend these ideas to more complex situations. There is, of course, the hope that professionals trained in classical methods of data analysis will attempt to acquaint themselves with the general linear hypothesis approach to data analysis.

Of greater concern to this reviewer is the apparent mistaken

notion that because the model specified unbiased estimates of regression coefficients can be obtained, the need to meet the assumptions underlying the traditional analysis of variance procedures is not important. On p. 70, they conclude: "In an approximate test the probability of making a Type I error is not exactly known, but the researcher often assumes that this probability is close to the level of significance." However, as is well known from the work of Box and Scheffé such assumptions should be restricted to the case where equal treatment groups are available. Such unqualified statements can lead the relatively unsophisticated in statistics to believe that anything goes if regression analysis is used.

The use of regression analysis with higher factorial and hierarchical designs with unequal sample sizes also requires some comment. When treatment group sizes are unequal differing orders of entering model effects into the regression equation will result in differing sums of squares. Also, the algorithm utilized in regression analysis may result in different conclusions. That is, some computer programs will report only the SS remaining after a previous effect has been accounted for and some programs will provide sums of squares corrected for the unbalance in the design.

In summary, this author feels that a source describing the general linear hypothesis approach to data analysis should be available to researchers with a behavioral or social science background. However, the enthusiastic approach provided by Kelly and his colleagues appears to only point out the good aspects and none of the problems associated with regression analysis. Joe Ward and Earl Jennings are presently at work on a book which is concerned with the same methods in data analysis; it would seem that a person who wants the last word on this subject should await this upcoming text before making a choice as to which single text on regression analysis he should have in his library.

JOHN L. WASIK
Center for Occupational Education
North Carolina State
University at Raleigh

Howard B. Lyman. *Test Scores and What They Mean*. (2nd ed.) Englewood Cliffs, N. J.: Prentice-Hall, 1971. Pp. viii + 200. \$6.95 and \$4.95 (paperback).

The second edition of *Test Scores and What They Mean* is unique in many ways. Like the first edition, it differs from the typical book devoted to educational and psychological testing in a number of important respects. Primary among these is the fact that it concentrates heavily on test scores and their meaning, including

only minor amounts of information about such topics as test construction, scoring, and administration. Although a variety of standardized tests—particularly those in the area of achievement and aptitude—are mentioned for illustrative purposes, the author avoids detailed descriptions of such instruments. In short, the title is reasonably descriptive.

Another noteworthy aspect of the book is that it is designed to assist practicing professionals who are not knowledgeable about testing, for example, school teachers, social workers, admissions counselors, and even psychiatrists and pediatricians. In addition, the author believes that the book is suitable for use in connection with college level courses in educational psychology, educational and psychological testing, and guidance.

The first two of its 12 chapters introduce the reader to modern testing practices and associated vocabulary. Following these is a highly abbreviated chapter devoted to test validity and reliability, and another concerning statistical methodology, the emphasis here being primarily on descriptive statistics. With the exception of the reliability material, these four chapters differ little from those found in the first edition.

The fifth chapter is a discussion of the test manual and is one of two new chapters in the second edition. The heart of the book is represented by the two chapters which follow, namely, one concerning derived scores and another concerning profiles. The former is easily the longest chapter in the book and the most technical. Various types of derived scores are classified in a rather elaborate manner and systematically described. This is no doubt one of the most complete chapters of its kind. In contrast, the chapter concerning profiles is shorter and replete with illustrations taken from common standardized tests.

The next three chapters are comparatively short, one consisting of only two pages. This arrangement is unusual, to say the least, since the theme of the three is embodied in the title of the first, namely, "Common Sense." The usual list of "do's and don'ts" of test score interpretation and the communication of such scores are listed, and the "don'ts" are often illustrated. These three chapters along with the summary chapter at the end of the book regularly give the reader the feeling that the author is painstakingly elaborating the obvious. On the other hand, repeated mention of common precautions to be taken in test score interpretation may well be in order for the kind of audience for which this book is intended.

Chapter Eleven is the second of the two new chapters added to the second edition. Eight criticisms of psychological testing are listed and the author's reaction to each is provided. Almost all of these bear directly upon test score interpretation and consequently this chapter is a vital addition to the book. Unfortunately, the

treatment of each criticism is extremely limited. This chapter could be greatly strengthened if a reasonable sampling of the many excellent references regarding these criticisms were added at appropriate points.

The appendices include a glossary of terms used in psychological testing, a bibliography of books, monographs, and audio-visual material concerning testing, and—most significant of all—a table for converting one derived-score system to another if the assumption of an underlying normal distribution is satisfied. This table is one of the unique contributions made by the book and is designed in a manner consistent with the classification scheme for derived scores used in Chapter Six.

It is easy to understand why this volume would be comparatively popular with practicing professionals who have had little or no formal instruction in psychological testing. Its level of difficulty is quite low in almost all instances. The author's approach is strictly "cookbook"; rarely does he attempt to treat a topic in depth. His style is informal, if not folksy. Certainly no other book in this field is laced so heavily with the use of the first person singular.

Considering the audience to which this book is directed and the nature of its emphasis, its relatively short length (200 pages) is a distinct blessing. Nevertheless, it is extremely difficult to avoid recommending that the author strengthen his presentation with a number of additions here and there. For instance, one wonders, why the discussion of test validity is so limited and not systematically cross-referred with other related sections of the book. Also, the use of annotated bibliographies at the end of each chapter would definitely improve the volume in that it would offer an easy route for an eager reader to acquire deeper views of any topic of particular interest to him. Finally, there are less noticeable oversights such as the failure to give proper attention to criterion-referenced tests, ipsative scores, many of the test publications edited by Buros, and the increasingly vital role played by computers in test score reporting and interpretation.

In summary, the book is a tidy and simple presentation of test scores and their interpretation, and in many ways is well designed for the uninitiated. All others will probably find little in the book to attract them, unless it is the chapter devoted to derived scores and the conversion table for them found in the appendix.

J. STANLEY AHMANN
Colorado State University

George G. Stern. *People in Context*. New York: John Wiley, 1970.
Pp xxvi + 402. \$13.95.

We live in a world of organizations in which individuals find meaning and expression in their lives chiefly through the ways in

which they identify with organizations and the ways in which those organizations make possible the realization of individual aspirations. A particularly crucial relationship between individual and organization is that between a student and his college or university. It is this relationship that George Stern explores in great detail in his book, *People in Context*. The nature of this relationship, according to Stern, is a dynamic one which requires the development of a technique which recognizes, as did Kurt Lewin, that person and environment be represented in common terms as complementary parts of an interaction.

The means that Stern and his collaborators at the University of Syracuse have chosen to explore this dynamic relationship are based on the Need-Press model developed by Henry A. Murray. The Need-Press model is a taxonomy of psychogenic needs which are related to environmental characteristics which match these psychogenic needs. Needs are identified as "characteristic spontaneous behaviors manifested by individuals in their life transactions." Press, as the complement of needs, is made up of "characteristic behaviors manifested by aggregates of individuals in their mutual interpersonal transactions." In the situations Dr. Stern investigates, the psychogenic needs are those of college students and the environmental characteristics are those of colleges and universities. The interaction of needs and press may produce congruence, growing out of favorable, compatible environmental circumstances matching and enhancing personal needs; or forms of dissonance which result from a poor match of personal needs and environmental characteristics. If it is possible to implement the model by developing effective measures of need and press and to demonstrate the ecological dynamics of the interactions, it may then be possible to influence both individual adjustment and collegiate environment so as to bring about productive educational experiences, and those tasks are what Dr. Stern is undertaking in this book.

Much of the book is devoted to an account of the development of the measuring instruments chosen by Stern and his colleagues to develop his ideas. The basis of these instruments, of course, is the Murray need catalog consisting of some 30 needs. The measuring instrument developed is the Activities Index (AI) derived from its original prototype first developed at Chicago in the early 1950's. The instrument used in the research reported in this book consists of 300 items distributed, 10 items each, over 30 scales based on the Murray needs. The environmental counterpart of the personal needs consists of a series of Environmental Indexes. The first and most important of these indexes and the one most fully reported in this research is the College Characteristics Index (CCI). The other Environmental Indexes are High School Characteristics Index (HSCI), Evening College Characteristics Index (ECCI), and Or-

ganizational Characteristics Index (OCI). This review will refer subsequently only to the College Characteristics Index (CCI). Both Indexes (AI and CCI) are self-administered questionnaires. The reporting of the development and validation of these instruments is extensive in this book. The Indexes themselves are printed in their entirety in one of the appendices.

The bulk of the book is taken up with the reporting of the characteristics of various college populations and various college and university environmental settings as measured by the Indexes. Extensive study of the characteristics of the Indexes and factor analyses of the Indexes are also reported. The factors derived are used to bring about deeper understanding of collegiate behavior. A set of first-order factors extracted from the AI and CCI scales produces some 12 factors for AI and 11 for the CCI. A second-order factor analysis of those factors demonstrated three major dimensions among the student personality factors. These three dimensions are (1) *Achievement Orientation* which includes factors Self-Assertion, Audacity-Timidity, Intellectual Interests, Motivation, Applied Interests; (2) *Dependency Needs* with factors Applied Interests, Constraint-Expressiveness negative, Diffidence-Egoism negative, Orderliness, Submissiveness, Timidity-Audacity negative, and Closeness; (3) *Emotional Expression*, containing factors Closeness, Sensuousness, Friendliness, Expressiveness-Constraint, Egoism-Diffidence, and Self-Assertion. A fourth Dimension, Educability, also appears but is of less magnitude than the other three.

The second-order factor analysis of environmental factors produced two major dimensions: (1) *Intellectual Climate* containing the factors Work-Play negative, Nonvocational Climate negative, Aspiration Level, Intellectual Climate, Student Dignity, Academic Climate, Academic Achievement, Self-Expression; and (2) *Non-intellectual Climate* with factors Self-Expression, Group Life, Academic Organization, Social Form, Play-Work, Vocational Climate.

The data on which the analyses of the instruments (AI and CCI) were based and which were used to answer the major questions to which the research was addressed were taken from the responses made by students to the two Indexes at a large number of colleges and universities throughout the United States. The major questions the research sought to answer were: (1) What are the major psychometric properties of the two instruments? (2) Can the factor scores be used to classify schools and student bodies? (3) Are the measures of personality and institutional press related to educational objectives and their achievement? (4) How do measures of environmental press for an institution as a whole relate to those for subcultures within the institution? (5) How is correspondence between personal needs and environmental press best expressed and quantified? The answers to these questions are extensively explored

and reported in the book. The nature of the pattern of needs and press revealed by the instruments and extended analysis of data produced by them has already been referred to above. The characteristics of special classes of schools and student bodies are elaborated in considerable detail and a reader interested in these phenomena has a rich collection of data to examine. There are, for example, clear and interesting differences among all of the basic classes of institutions investigated: Independent Liberal Arts, Denominational Liberal Arts, University Affiliated Liberal Arts, Business Administration, Engineering and Teacher Training. The differences are quite striking, for example, between the Independent, Denominational, and University Liberal Arts Colleges. The Independent Liberal Arts are markedly higher on Intellectual Climate and markedly lower than the others on Non-Intellectual Climate. On personality factors there are also quite striking differences between male and female students. The difference is greatest for Achievement Orientation, but is also marked for a number of factors in the dimensions Dependency Needs and Emotional Expression. As may be expected, a vast number of comparisons between various classes of schools and student categories can be made when one considers the large number of students and institutions from which data were gathered in this study. For the academic reader in particular these comparisons are fascinating and illuminating. This review of course can do no more than hint at the studies and comparisons presented in the book. In addition to the studies already briefly described, there are descriptions and analyses of denominational colleges and universities, of three particular institutions, of the variability among schools within a large university (Syracuse) and of various college climates.

Since this book, in its stated purposes and in the means it describes for achieving those purposes, seeks to present to its readers in the education profession some fundamental understanding of colleges and universities and some ways of making those institutions more effective, it is fair to judge it on its contribution to those goals. Technically the work presented in this book is a most substantial achievement. The work that went into the development of these scales and their use in the measurement and description of collegiate behavior and settings is little short of prodigious. The design of the studies and the work involved in carrying them out is technically competent at the highest level. The book is well documented and referenced at all points and the professional scholar in the field will find it extremely useful. For the researcher who may wish to use some of the instruments and techniques in his own institution the enormous amount of data and technique presented make it a marvelously useful book.

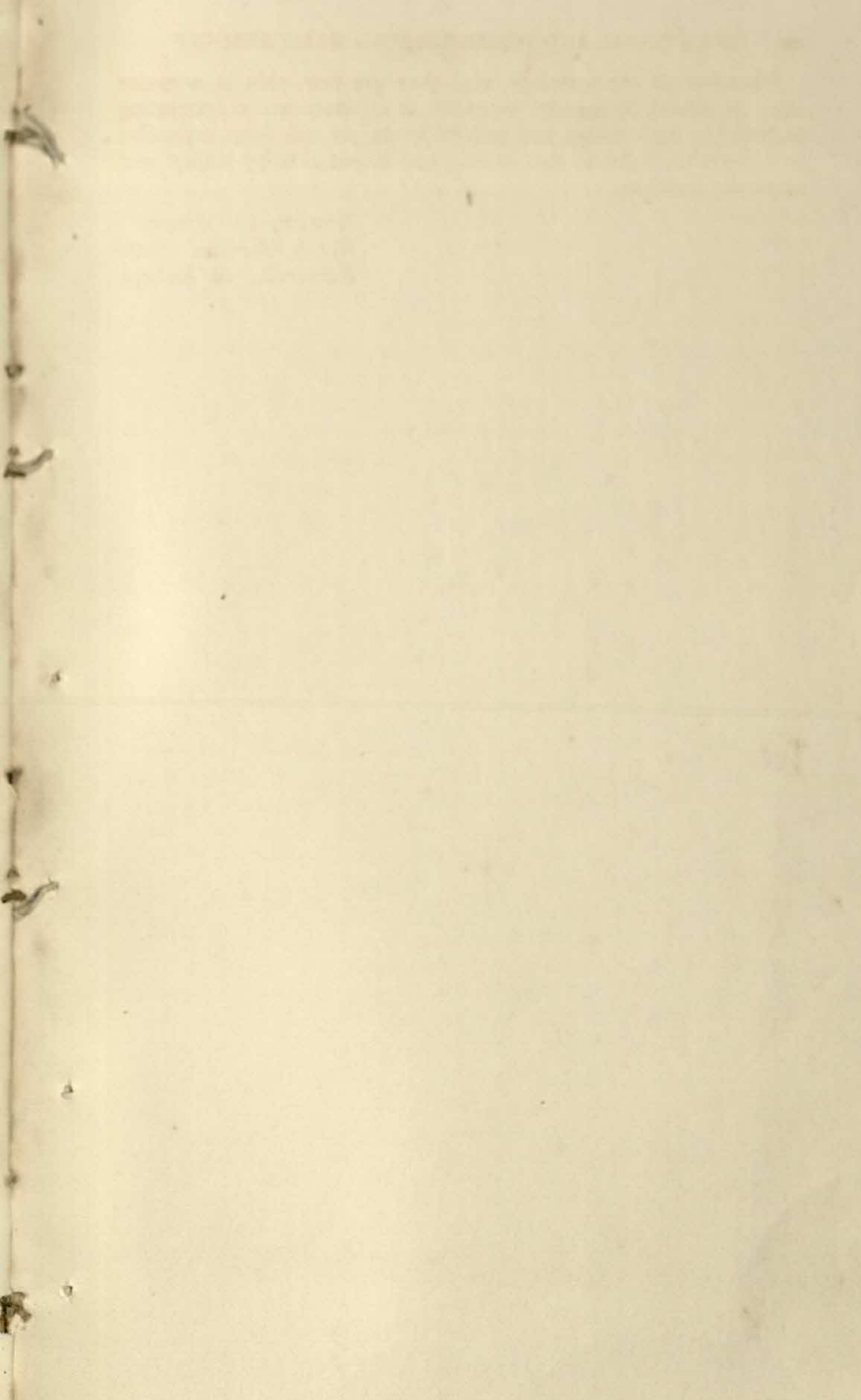
Conceptually it seems to this reviewer that the book has some limitations. The techniques used are certainly effective for describing institutions and people. The device of using a single set of concepts based on a theory of personality to develop and describe the relationship between person and setting is unique and effective. But the technique results in a static portrayal of both students and schools and the results of their interaction. The description of both students and institutions are limited to the accumulated scores on instruments on which the responses made by students are confined to yes and no answers to rather narrowly drawn questions. Even though this is a useful method for many purposes it has qualities of barrenness when it is the only method used for describing and understanding the dynamics of human behavior as it occurs in such fascinating settings as colleges and universities.

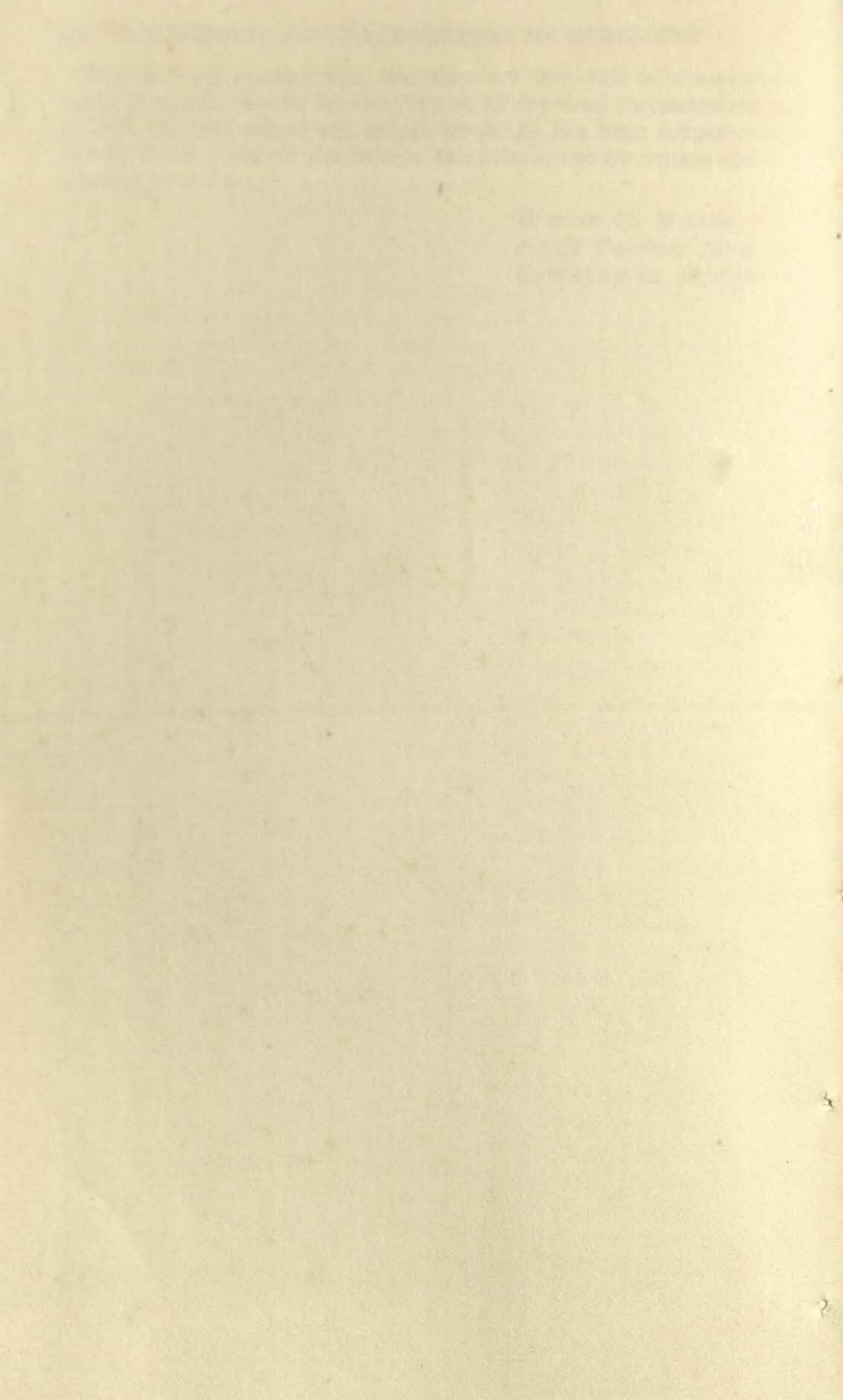
The descriptions and conclusions in the book, though brilliant and fascinating, rest solely on the accumulation and manipulation of psychometric data. The report is almost wholly unrelieved by the reporting of incidents and cases which illustrate the uses of this technique in learning about individual college students and their functioning. There is only one case study reported—one, incidentally, which demonstrates rather convincingly the power of the techniques to predict and understand the behavior of an individual student. More such studies would have helped greatly to demonstrate the usefulness of the work. Some accounts of social interactions of various groups in various settings might also have been illuminating. The techniques used by Stern, even though masterfully done, as is the case here, have their limitations when it comes to understanding the dynamics of human behavior.

There are some important uses to which these ideas and instruments may be put. There has always been a rather general belief in this country that students and colleges are basically very much alike. Though this idea has been often suspect, there has been little in the way of appropriate conceptual schemes and adequate measuring instruments to give substance to the suspicions and to provide a real means to think and talk about such differences. It has been particularly true of faculties that after being trained in high intellectual colleges and in the rarified atmosphere of graduate schools they have been blind to the particular qualities of their students or have assumed them to be inferior versions of their classmates or the students they taught while graduate assistants. For educational reasons it is important that these faculties come to understand their students and the psychological climates in which they operate. The Stern instruments and the Need-Press conceptual scheme offer a superb way for them to do so.

Whatever its shortcomings, and they are few, this is a major book. It should be read by everyone at all involved in counseling at both the high school and college levels. As has been suggested as well above, it should also be read and attended to by college and university faculties.

HOWARD G. MILLER
*North Carolina State
University at Raleigh*





EDUCATIONAL and
PSYCHOLOGICAL



MEASUREMENT

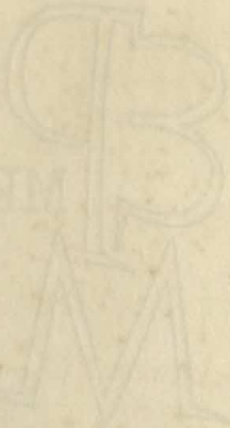
Editor: W. Scott Gehman

Managing Editor: Geraldine R. Thomas

BOARD OF COOPERATING EDITORS

- DOROTHY C. ADKINS, *University of Hawaii*
LEWIS R. AIKEN, JR., *Guilford College*
HAROLD P. BECHTOLDT, *The University of Iowa*
WILLIAM V. CLEMANS, *Science Research Associates, Inc.*
LOUIS D. COHEN, *University of Florida*
JUNIUS A. DAVIS, *Educational Testing Service*
HAROLD A. EDGERTON, *Performance Research, Inc.*
MAX D. ENGELHART, *Duke University*
GENE V GLASS, *University of Colorado*
E. B. GREENE, *Chrysler Corporation (Retired)*
J. P. GUILFORD, *University of Southern California, Los Angeles*
JOHN A. HORNADAY, *Babson College*
JOHN E. HORROCKS, *The Ohio State University*
CYRIL J. HOYT, *University of Minnesota*
MILTON D. JACOBSON, *University of Virginia*
JOSEPH C. JOHNSON II, *Greenwich Public Schools*
WILLIAM G. KATZENMEYER, *Duke University*
E. F. LINDQUIST, *State University of Iowa*
FREDERIC M. LORD, *Educational Testing Service*
ARDIE LUBIN, *Naval Medical Neuropsychiatric Research Unit, San Diego*
LOUIS L. MCQUITT, *University of Miami, Coral Gables*
WILLIAM B. MICHAEL, *University of Southern California, Los Angeles*
HOWARD G. MILLER, *North Carolina State University at Raleigh*
ELLIS B. PAGE, *The University of Connecticut*
NAMBURY S. RAJU, *Science Research Associates, Inc.*
BEN H. ROMINE, JR., *University of North Carolina at Charlotte*
KENDON SMITH, *The University of North Carolina at Greensboro*
THELMA G. THURSTONE, *University of North Carolina at Chapel Hill*
HERBERT A. TOOPS, *The Ohio State University*
WILLARD G. WARRINGTON, *Michigan State University*
JOHN E. WILLIAMS, *Wake Forest University*
E. G. WILLIAMSON, *University of Minnesota*

VOLUME THIRTY-ONE, NUMBER THREE, AUTUMN 1971



MEASUREMENT

MEASUREMENT

MEASUREMENT

A FACTOR ANALYTIC INTERPRETATION STRATEGY

MARGARET L. HARRIS

Research and Development Center for Cognitive Learning
University of Wisconsin

CHESTER W. HARRIS¹

Psychometric Laboratory
University of Wisconsin

THE purpose of this paper is to illustrate the use of a strategy for determining the common factors in a set of data. C. Harris (1967) suggested using several different computing algorithms for the initial solution, obtaining derived solutions, both orthogonal and oblique, comparing the results, and regarding as the important substantive findings those factors that are robust with respect to method. This paper illustrates a way of comparing the results.

The factor results used for this illustration of a factor analytic interpretation strategy are the reanalyses, by seven different solutions, of the data from nine of the Guilford studies as reported by C. Harris (1967). The initial component and factor methods used are Incomplete Principal Component (Hotelling, 1933), Alpha (Kaiser and Caffrey, 1965), a Jöreskog method (1963, 1967), and Harris R-S² (1962). The Jöreskog method used for Matrices 08 and 23 is his Unrestricted Maximum Likelihood Factor Analysis (UMLFA) procedure (1967) using a critical value of .05; Jöres-

¹ Now at the Graduate School of Education, University of California, Santa Barbara.

The research reported herein was performed, in part, pursuant to a contract with the United States Office of Education, Department of Health, Education, and Welfare, under the provisions of the Cooperative Research Program. The opinions expressed in this publication do not necessarily reflect the position or policy of the Office of Education and no official endorsement by the Office of Education should be inferred. Center No. C-03/Contract OE 5-10-154.

kog's early procedure (1963) was used for the other seven matrices. These four methods provide a component solution (Incomplete Principal Component), a factor solution with a statistical basis (Jöreskog, 1963 or UMLFA), and two factor solutions with a psychometric basis—one for a minimum number of factors (Alpha) and one for a maximum number of factors (Harris R-S²). It may be noted that these three factor methods are scale-free. Derived orthogonal solutions were obtained for each of the four initial solutions using the Kaiser normal varimax² procedure (1958) and derived oblique solutions were obtained for the first three initial solutions using the Harris-Kaiser independent cluster solution (1964). An oblique solution was not obtained for the Harris R-S² method since it would have certain correspondences to the oblique solution obtained from the Jöreskog (1963) method.

The nine Guilford matrices that were reanalyzed are:

- 08 Creative thinking
- 09 Evaluative abilities
- 12 Planning
- 14 General reasoning
- 16 Reasoning, creativity, and evaluation
(Subdivided into three—16A, 16B, and 16C)
- 22 Problem-solving abilities
- 23 Cognition and convergent production

The number of initial factors (components) obtained for each of the four methods for each of the nine matrices is given in Table 1. Also included in this table is the number of common factors (components) obtained for each of the seven derived solutions. A common factor (component) is defined as one having at least two variables with coefficients greater than .30 (absolute). All of the common factors are utilized for the interpretation strategy illustrated in this paper; thus, all of the variables with values greater than .30 (absolute) on one or more common factors appear in the tables. Note that Guilford used a coefficient of .30 (absolute) as a critical value in interpreting his derived orthogonal factors.

² Guilford and Hoepfner (1969) have compared varimax rotations with rotations to theoretical targets and essentially rejected the former as not giving meaningful results. It seems likely that they would find our results unsatisfactory since these do not reproduce the Structure of Intellect model in detail but instead suggest alternative interpretations.

TABLE 1

Numbers of Initial and Derived Common Factors for the Various Methods

Matrix	Factor Method	Initial Factors	Orthogonal Common Factors	Oblique Common Factors
08	Incomplete Principal Component	14	13	14
	Alpha	14	11	13
	UMLFA	19	10	14
	Harris R-S ^a	28	10	
09	Incomplete Principal Component	15	12	14
	Alpha	15	13	13
	Jöreskog	*		
	Harris R-S ^a	39	11	
12	Incomplete Principal Component	13	12	13
	Alpha	13	10	12
	Jöreskog	7	7	7
	Harris R-S ^a	30	7	
14	Incomplete Principal Component	6	6	6
	Alpha	6	6	6
	Jöreskog	4	4	4
	Harris R-S ^a	13	7	
16A	Incomplete Principal Component	6	5	6
	Alpha	6	5	5
	Jöreskog	4	4	4
	Harris R-S ^a	16	7	
16B	Incomplete Principal Component	6	6	5
	Alpha	6	6	5
	Jöreskog	4	4	4
	Harris R-S ^a	14	7	
16C	Incomplete Principal Component	6	6	6
	Alpha	6	6	6
	Jöreskog	6	5	6
	Harris R-S ^a	14	6	
22	Incomplete Principal Component	12	11	11
	Alpha	^b		
	Jöreskog	7	7	7
	Harris R-S ^a	24	8	
23	Incomplete Principal Component	5	5	5
	Alpha	5	5	5
	UMLFA	5	5	5
	Harris R-S ^a	17	6	

^a Went to $p - 1$ factors.^b Did not converge.

The procedure involves attempting to find the common factors (components) that are similar over solutions. This was done by starting with a derived orthogonal component from the Incomplete Principal Component initial method. The reason for this choice is that this solution tends to include more variables with coefficients greater than .30 on a particular component than any of the other solutions. Then for each other derived orthogonal solution and for each derived oblique solution, a common factor was searched for that seemed to be similar to the component selected, particularly with respect to the large coefficients.

The next step involved determining those factors (components) that are robust with respect to method—factors which tend to include the same variables across methods. A variable was considered relevant to a factor if it had a coefficient greater than .30 (absolute) on that factor. A comparable common factor (CCF) was defined as one having two or more of the same relevant variables on at least five of the seven derived factors (components). This means that a comparable common factor is defined by more than two different initial solutions and by both orthogonal and oblique rotations. Thus, no one initial method can account for a variable's rejection and no one derived method can account for a variable's acceptance on a comparable common factor. Note that for the two matrices for which one of the initial solutions was not available, Matrix 09 and Matrix 22, a comparable common factor is defined as one having two or more of the same relevant variables on at least four of the five solutions.

Two other types of factors may be found. A comparable specific factor (CSF) is defined as one having only one (the same) relevant variable on at least five of the solutions. A noncomparable factor (NCF) is defined as one not having any one or more of the same relevant variables on at least five of the solutions.

Table 2 contains the number of comparable common factors, comparable specific factors, and noncomparable factors for each of the nine matrices. The number of common factors obtained by Guilford for each matrix is also given in Table 2.

The two matrices chosen as illustrations of the strategy for this paper are 23 and 08.³ Matrix 23 was chosen to illustrate the fairly

³ The application of the interpretation strategy proposed in this paper to all nine of the Guilford studies can be found in Harris and Harris (1970).

TABLE 2
Number of Factors for Each Matrix

Matrix	Reanalyses			Guilford
	Comparable Common Factors	Comparable Specific Factors	Noncomparable Factors	Common Factors
08	10	0	8	15
09	10	1	10	14
12	7	2	4	14
14	6	0	1	9
16A	4	1	5	11
16B	5	0	3	9
16C	5	0	5	10
22	7	0	7	13
23	5	0	1	13

close agreement across methods that can be secured among various factor solutions. Matrix 08 was chosen as a matrix for which the various factor solutions are in least agreement. Of the nine matrices studied, the results for 08 and 09 seemed to be the most discrepant across the seven derived solutions. Of these two, Matrix 08 was chosen for presentation here because one initial factor method was not available for Matrix 09. For 08 the various solutions agree in part but for some of the factors the results are quite diverse. Table 3 contains the results for Matrix 23 and Table 4 the results for Matrix 08. The relevant variables are in capital letters and the nonrelevant variables (noise?) are in small letters. The order of the factors in the tables is arbitrary within each of the three types of factors (CCFs, CSFs, and NCFs). Guilford's results are presented in each table with the factors of the reanalyses with which they seem to agree most closely.

For Matrix 23 the factors are rather robust over solutions. There are five comparable common factors for the 30 variables in this matrix and one noncomparable factor. This is in contrast to the 13 common factors obtained by Guilford.

As shown in Table 4, the results for Matrix 08 are not as robust over solutions as they are for Matrix 23; the results from the various solutions are comparable (in the sense defined for this strategy) for some factors but not for others. It should be pointed out here that for both 10 and 12 factors the UMLFA method yielded an improper solution since the unique variance for variable num-

	Reanalyses								Guilford
	Orthogonal				Oblique				
	I	II	III	IV	I	II	III		
COMPARABLE COMMON FACTOR 1									
9 LIMITED SUMS	56	43	45	44	48	42	55	E	37
12 NUMBER RELATIONS	46	38	48	39	32	36	57		
14 NUMERICAL OPERATIONS	59	39	45	58	54	39	59		51
16 OPERATIONS SEQUENCE	50	45	53	43	37	42	61		
19 PICTURE-GROUP NAMING	-66	-51	-41		-78	-61	-67		
COMPARABLE COMMON FACTOR 2									
2 CAMOUFLAGED WORDS	52	41	39	38	47	40	47	A	
12 NUMBER RELATIONS	37	37	37		34	43	40		
17 ORDERING I	77	62	52	58	87	79	74		36
23 VERBAL COMPREHENSION	64	49	57	48	62	50	81		51
30 WORD TRANSFORMATIONS	44	39	43	46	33	33	49		
1 Alterations			32						
20 Seeing Trends II	38	34	42				43		32
22 Symbol Grouping					-35		-32		
26 Word Fluency					-31				
27 Word Groups	43	38	46				49		41

* Decimals have been omitted.

Key to Factor Solutions of Reanalyses:

- I Incomplete Principal Component
- II Alpha
- III UMLFA
- IV Harris R-S²

Key to Guilford Factors:

- E Numerical Facility
- A Verbal Comprehension

TABLE 3*(Continued)

	Reanalyses										Guilford					
	Orthogonal					Oblique					B	D	G	H	L	M
	I	II	III	IV	I	II	III									
COMPARABLE COMMON FACTOR 3																
3 CIRCLE REASONING	54	45	46	48	65	52	55							40		
7 LETTER GROUPING	53	55	54	51	49	48	46							40		
8 LETTER TRIANGLE	69	65	62	68	86	80	72							42		
13 NUMBER SERIES CORRECTION	40	40	36	38	37	34						44		31		
16 OPERATIONS SEQUENCE	55	58	56	56	53	48										52
18 PICTURE ARRANGEMENT	56	46	45	42	74	62	61						55			
20 SEEING TRENDS II	54	49	48	44	60	54	50									
22 SYMBOL GROUPING	61	53	52	45	75	66	64									
24 WORD CHANGES	64	65	63	64	69	67	56							35		35 30
27 WORD GROUPS	47	43	43	42	48	41	35									36 49
28 WORD PATTERNS	43	40	39	31	42	38	39									
29 WORD RELATIONS	57	56	56	55	59	54	52									50
1 Alterations	36	36	36	36												
6 Letter Analogies	32	35	33													
9 Limited Sums	32	36	31	34								37		31	49	
10 Number Classification																
12 Number Relations	35	37	33	38								45		36		
17 Ordering I																
21 Ship Destination			56				64					50		C	K	
COMPARABLE COMMON FACTOR 4																
5 FOUR-LETTER WORDS																
7 LETTER GROUPING	45	36	33		44	33	31									
10 NUMBER CLASSIFICATION	40	38	36	33										37		
11 NUMBER-GROUP NAMING	71	56	57	63	75	61	66							41		
12 NUMBER RELATIONS	83	77	79	66	93	91	94							45		
13 NUMBER SERIES CORRECTION	40	42	35	37	31									44		
6 Letter Analogies	40	34	33	35	35									32		
19 Picture-Group Naming	32															
27 Word Groups	37				36		38							50		32

L Cognition of Symbolic Implications
M Convergent Production of Symbolic

Key to Guilford Factors:
B General Reasoning
D Ordering

TABLE 3 (Continued)

	Reanalyses								Guilford		
	Orthogonal				Oblique						
	I	II	III	IV	I	II	III				
COMPARABLE COMMON FACTOR 5											
1 ALTERATIONS	48	48	43	47	41	40		F	I	J	
2 CAMOUFLAGED WORDS	43	42	39	44	40	35		31	44	32	
4 DISEMVOELED WORDS	78	76	77	73	91	96	85	41	36	53	
5 FOUR-LETTER WORDS	43	38	41	40	44	36	36	47	49	49	
15 OMELET TEST	73	70	72	71	82	82	75	31	50	52	
25 WORD COMBINATIONS	65	60	60	63	69	64	56				
26 WORD FLUENCY	70	55	52	54	89	74	61				
28 WORD PATTERNS	44	37	39	34	41	34	35				
30 WORD TRANSFORMATIONS	55	55	50	54	53	49	33				
6 Letter Analogies			31								
7 Letter Grouping			38								
12 Number Relations	38	39	38	37							
18 Picture Arrangement				32							
27 Word Groups		32			-35						
29 Word Relations	41	42	36	37							
NONCOMPARABLE FACTOR 6											
20 Seeing Trends II											
27 Word Groups											

Key to Guilford Factors:

- F Word Fluency
 I Symbolic Redefinition
 J Cognition of Symbolic Units

TABLE 4
Factor Results for Matrix 08^a

	Reanalyses							Guilford	
	Orthogonal				Oblique				
	I	II	III	IV	I	II	III		
COMPRARABLE COMMON FACTOR 1								D	
35 PUNCHED HOLES	57	49	50	52	48	37	34	45	
48 PRACTICALJUDGMENT	60	47	38	46	68	56	37	32	
51 MECHANICAL PRIN- CIPLES	80	71	78	69	86	78	80	54	
52 ARITHMETIC REASON- ING	45	44	51	49	36	33		38	
16 Match Problems	41	38		34					
34 Word Matrices	31								
COMPARABLE COMMON FACTOR 2								C	F
36 MUTILATED WORDS	40	36	38		33		50	35	
37 STREET GESTALT COMPLETION	70	62	63	59	71	70	64	37	44
38 PERCEPTUAL SPEED	64	58	54	57	65	47		56	
41 UNUSUAL DETAILS	34	33	33	31			34		
42 PENETRATION OF CAMOUFLAGE	76	67	68	67	81	72	55	45	40
47 SPATIAL ORIENTATION (PART I)	60	54	50	53	59	44		47	
35 Punched Holes		32							

^a Decimals have been omitted.

Key to Factor Solutions of Reanalyses:

I Incomplete Principal Component

II Alpha

III UMLFA

IV Harris R-S²

Key to Guilford Factors:

D Visualization

C Perceptual Speed

F Closure

TABLE 4 (Continued)

	Reanalyses							Guilford
	Orthogonal				Oblique			
	I	II	III	IV	I	II	III	
COMPARABLE COMMON FACTOR 3								B
49 NUMERICAL OPERATIONS (PART I)	83	76	76	74	93	90	80	72
50 NUMERICAL OPERATIONS (PART II)	78	70	77	73	83	79	82	73
52 ARITHMETIC REASONING	50	45	43	43	43	38		49
1 Sentence Analysis					-31			
44 Ship Destination	33							
47 Spatial Orientation (Part I)	37				35			37
COMPARABLE COMMON FACTOR 4								A E
1 SENTENCE ANALYSIS	39	32	33		43	37	51	
2 PARAGRAPH ANALYSIS	49	35	35		60	47	41	
27 SENTENCE GESTALT (OMISSIONS)	-60	-54		-47	-76	-75		53
33 SENTENCE SYNTHESIS	65	64	65	62	64	63	67	53
43 VOCABULARY	71	68	70	70	74	74	61	65
46 INFERENCE TEST	67	63	64	58	68	64	73	47
53 SENTENCE GESTALT	42	46	48	47	32			43
11 Number Associations (Uncommonness)		31						
14 Circle Square I			34					31
15 Circle Square II	40	43	46	37			32	
17 Sign Changes	34	35	38					36
18 Implied Uses	32	37	42	39			38	
21 Associations II			31					
28 Word Transformation			32					
32 Concept Synthesis	32	37	39	40				35
34 Word Matrices		35	40	32				
44 Ship Destination	36	35	38					38
51 Mechanical Principles							32	42
52 Arithmetic Reasoning	34	39	42	32				33
								35

Key to Guilford Factors:

B Numerical Facility

A Verbal Comprehension

E General Reasoning

TABLE 4 (Continued)

	Reanalyses							Guilford
	Orthogonal				Oblique			
	I	II	III	IV	I	II	III	
COMPARABLE COMMON FACTOR 5								N
24 APPARATUS TEST	69	59	60	61	70	67	60	59
25 SOCIAL INSTITUTIONS (DIRECT)	80	67	75	66	90	84	83	70
13 Consequences (Remote)	37				31			
22 Unusual Uses	32							
41 Unusual Details	31							
44 Ship Destination	31				34			
COMPARABLE COMMON FACTOR 6								G H
28 WORD TRANSFORMA- TION	72	59	52	58	75	70	57	52 32
40 DISARRANGED WORDS	72	54	57	53	79	62	32	38
53 SENTENCE GESTALT	55	49	48	43	44	45	92	56
11 Number Associations (Uncommonness)								33
14 Circle Square I								36
15 Circle Square II	38	31						44
27 Sentence Gestalt (Omissions)								34
36 Mutilated Words	31							37
39 Controlled Associations	31		32					46
COMPARABLE COMMON FACTOR 7								K
16 MATCH PROBLEMS	44	32	32		47	43	45	37
45 SYMBOL MANIPULA- TION	62	40	44		65	49	36	32
17 Sign Changes								
23 F-Test					-36			
35 Punched Holes							33	
38 Perceptual Speed	35		32		35	45		
40 Disarranged Words							35	
43 Vocabulary					-31	-37	-43	
53 Sentence Gestalt						-31		

Key to Guilford Factors:

N Sensitivity to Problems

H Associational Fluency

G Word Fluency

K Adaptive Flexibility

TABLE 4 (Continued)

	Reanalyses							Guilford
	Orthogonal				Oblique			
	I	II	III	IV	I	II	III	
COMPARABLE COMMON FACTOR 8								M
29 GESTALT TRANSFOR- MATION	61	50	48	38	56	39	53	37
30 PICTURE GESTALT	67	48	45	52	79	69	46	
19 Quick Responses (Uncommonness)							-37	
23 F-Test					-36			
31 Object Synthesis								31
41 Unusual Details					35	32		
48 Practical Judgment								31
COMPARABLE COMMON FACTOR 9								I
5 IMPOSSIBILITIES	50	35		44	45	39		39
6 PLOT TITLES (LOW QUALITY)	70	58		57	82	77		59
8 COMMON SITUATIONS	75	50		50	69	68		55
9 BRICK USES (FLUENCY)	74	48		49	72	66		54
12 CONSEQUENCES TEST (LOW QUALITY)	71	64		65	86	80		55
1 Sentence Analysis	38				32			
3 Figure Analysis	43				35			
4 Figure Concepts (Uncommonness)	33							
13 Consequences Test (Remoteness)	36							
22 Unusual Uses	44							
24 Apparatus Test	35							
31 Object Synthesis	34							
39 Controlled Associations	52							
41 Unusual Details	32							

Key to the Guilford Factors:

M Redefinition

I Ideational Fluency

TABLE 4 (Continued)

	Reanalyses							Guilford	
	Orthogonal				Oblique			J	L
	I	II	III	IV	I	II	III		
COMPARABLE COMMON FACTOR 10									
10 BRICK USES (FLEXIBILITY)	55	52	50	53	61	58	56		43
18 IMPLIED USES	52	38	34	33	65	60		31	39
22 UNUSUAL USES	39	67	69	63	37	50	64		
39 CONTROLLED ASSOCIATIONS	41	50	47		40	46			
1 Sentence Analysis	31	35	39	35					
3 Figure Analysis		47	47	44			43		
4 Figure Concepts (Uncommonness)		53	51	55			49	32	
5 Impossibilities		46	53	41			41	31	
6 Plot Titles (Low Quality)			34						
7 Plot Titles (Cleverness)		44	50	42			45	55	
8 Common Situations		57	68	54			67	31	33
9 Brick Uses (Fluency)		52	63	49			67		
11 Number Associations (Uncommonness)		51	43	44			31		
13 Consequences Test (Remoteness)		62	65	65			63	42	33
19 Quick Responses (Uncommonness)		33	32	34				49	
20 Associations I		45	43	38					
23 F-Test			35						
24 Apparatus Test		33	40						
25 Social Institutions (Direct)			35						
31 Object Synthesis		32	39				39		
32 Concept Synthesis	34				45	31			
34 Word Matrices	44				64	53			
37 Street Gestalt Completion					42				
38 Perceptual Speed						-34			
41 Unusual Details			35						

Key to the Guilford Factors:

J Originality

L Spontaneous Flexibility

TABLE 4 (Continued)

	Reanalyses							Guilford
	Orthogonal				Oblique			
	I	II	III	IV	I	II	III	
NONCOMPARABLE FACTOR 11								
14 Circle Square I					-31			O
24 Apparatus Test		34	34				34	34
26 Social Institutions (Indirect)		63	71		84		74	45
NONCOMPARABLE FACTOR 12								
38 Perceptual Speed							63	
47 Spatial Orientations (Part I)							53	
NONCOMPARABLE FACTOR 13								
40 Disarranged Words							-35	
52 Arithmetic Reasoning							67	
NONCOMPARABLE FACTOR 14								
4 Figure Concepts (Uncommonness)	31							
6 Plot Titles (Low Quality)					-39	-33	-102	
7 Plot Titles (Cleverness)	70				84	71	34	
13 Consequences Test (Remoteness)	39				34			
19 Quick Responses (Uncommonness)	65				63	31		
48 Practical Judgment					36			
NONCOMPARABLE FACTOR 15								
14 Circle Square I				41				
15 Circle Square II				42				

Key to Guilford Factor:
O, "Doublet"

TABLE 4 (Continued)

	Reanalyses							Guilford
	Orthogonal				Oblique			
	I	II	III	IV	I	II	III	
NONCOMPARABLE FACTOR 16								
17 Sign Changes	47				50	34		
23 F-Test	-34							
29 Gestalt Transformation					33			
31 Object Synthesis	-56				-62	-46		
32 Concept Synthesis	35							
36 Mutilated Words	39				39			
NONCOMPARABLE FACTOR 17								
4 Figure Concepts (Uncommonness)	31				34			
11 Number Associations (Uncommonness)	42				46	34		
12 Consequences Test (Low Quality)					-35			
19 Quick Responses (Uncommonness)					42	47		
20 Associations I	64				75	55		
23 F-Test					33	31		
29 Gestalt Transformation					-32	-37		
NONCOMPARABLE FACTOR 18								
11 Number Associations (Uncommonness)								35
18 Implied Uses								36
39 Controlled Associations								73

ber 27, Sentence Gestalt (Omissions), was equal to or less than .02. Jöreskog suggests partialling out any variables that have a unique variance that is essentially zero ($\leq .02$). It was decided, instead, to remove this variable from the intercorrelation matrix. The solution given here for UMLFA is for 15 factors for 52 variables, with variable number 27 omitted. There are 10 comparable common factors for the 53 variables in Matrix 08 and eight noncomparable factors. Guilford obtained 15 common factors for this set of data.

The results of the application of our factor analytic interpretation strategy to the remaining seven matrices are summarized only

in this paper. As mentioned earlier the seven derived solutions seemed to be very similar for Matrix 23. They are most discrepant for Matrices 08 and 09. The results seem to be fairly similar for Matrices 14 and 16B. For Matrices 12, 16A, 16C, and 22 there is some close agreement and some diversity. The comparable common factors of Matrices 09 and 22 seem to have relatively few relevant variables.

In general, the number of comparable common factors is similar to the smallest number of common factors in the derived solutions of the reanalyses. For one matrix (09) the number of CCFs is one less than the smallest number of common factors obtained for any one derived solution. For six of the matrices (08, 12, 16A, 16C, 22, and 23) the number of CCFs is equal to the smallest number of common factors obtained for any one or more derived solutions. The number of CCFs is greater than the smallest number of common factors for a single derived solution for two of the matrices (14 and 16B).

The number of comparable common factors for the data in any one of the matrices is always considerably fewer than the number of common factors obtained by Guilford. In general, a few of the CCFs agree rather closely with common factors obtained by Guil-

TABLE 5
Intercorrelations of Oblique Factors for Matrix 23^a

Comparable Common Factor	1	2	3	4
2-I ^b	04			
II	05			
III	54			
3-I	28	48		
II	35	60		
III	69	70		
4-I	25	29	48	
II	37	39	56	
III	58	55	55	
5-I	23	52	61	44
II	32	66	66	53
III	57	70	59	49

^a Decimals have been omitted.

^b Key to initial solutions:

- I Incomplete Principal Component
- II Alpha
- III UMLFA

ford. In many instances two or more of his common factors coalesce into one comparable common factor.

For all of the initial methods, the derived oblique solutions tend to drop variables with small coefficients from the common factors. Thus, more variables would be relevant to a comparable common factor, but with small coefficients, if only derived orthogonal solutions were used. Two good examples of this can be seen in CCF 3 of Matrix 23 (Table 3) and CCF 4 of Matrix 08 (Table 4).

The intercorrelations of the oblique factors are given, by initial method, in Table 5 for Matrix 23 and in Table 6 for Matrix 08.

TABLE 6
Intercorrelations of Oblique Factors for Matrix 08^a

Comparable Common Factor	1	2	3	4	5	6	7	8	9
2-I ^b	31								
II	39								
III	29								
3-I	35	32							
II	45	40							
III	30	25							
4-I	34	25	43						
II	45	34	53						
III	44	41	44						
5-I	02	12	15	15					
II	05	17	22	23					
III	-06	16	09	13					
6-I	23	34	44	46	05				
II	31	44	57	59	07				
III	31	47	53	63	11				
7-I	38	25	36	33	01	28			
II	57	42	53	47	05	47			
III	46	34	44	36	-01	36			
8-I	28	38	24	39	16	28	24		
II	36	54	34	51	28	40	38		
III	38	37	22	42	-02	32	30		
9-I	-01	11	11	12	53	07	05	20	
II	-01	15	16	19	44	10	09	33	
III									
10-I	29	31	30	49	31	35	28	43	40
II	33	34	36	58	42	44	39	57	54
III	07	25	09	32	42	24	13	03	

^a Decimals have been omitted.

^b Key to initial solutions:

- I Incomplete Principal Component
- II Alpha
- III UMLFA

These are included as an illustration of the possible comparability in some cases and diversity in other cases of the correlations of the derived oblique factors from the various initial methods that are included on the same CCF.

A strategy for determining comparable common factors in a given set of data has been illustrated. For one matrix, considerable agreement among the several derived solutions studied was demonstrated, but these results did not reproduce the ones secured initially by Guilford. For Matrix 23 this study offers a possible interpretation that does not support in detail Guilford's Structure of Intellect Model. For the other matrix, there was only a limited consistency among the derived solutions. The domain of creative thinking as defined by Matrix 08 appears to be unclear.

For future studies we would recommend obtaining both derived orthogonal and derived oblique solutions for each of these initial factor methods—Alpha, Harris R-S², and Unrestricted Maximum Likelihood Factor Analysis. A comparable common factor could then be defined as one having two or more of the same relevant variables on at least four of the six derived factors.

REFERENCES

- Guilford, J. P. and Hoepfner, Ralph. Comparisons of varimax rotations with rotations to theoretical targets. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 3-22.
- Harris, Chester W. Some Rao-Guttman relationships. *Psychometrika*, 1962, 27, 247-263.
- Harris, Chester W. On factors and factor scores. *Psychometrika*, 1967, 32, 363-379.
- Harris, Chester W. and Kaiser, Henry F. Oblique factor analytic solutions by orthogonal transformations. *Psychometrika*, 1964, 29, 347-362.
- Harris, Margaret L. and Harris, Chester W. A factor analytic interpretation strategy. Technical Report No. 115. Wisconsin Research and Development Center for Cognitive Learning, The University of Wisconsin, Madison, March 1970.
- Hotelling, Harold. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 1933, 24, 417-441, 498-520.
- Jöreskog, Karl G. *Statistical estimation in factor analysis*. Stockholm: Almqvist and Wiksell, 1963.
- Jöreskog, Karl G. Some contributions to maximum likelihood factor analysis. *Psychometrika*, 1967, 32, 443-482.
- Kaiser, Henry F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 1958, 23, 187-200.
- Kaiser, Henry F. and Caffrey, John. Alpha factor analysis. *Psychometrika*, 1965, 30, 1-14.

A COMPARATIVE STUDY OF SOME SELECTED METHODS OF PATTERN ANALYSIS^{1,2}

LOUIS L. McQUITTY

University of Miami
Coral Gables, Florida

THIS paper compares selected methods of pattern analysis by the author and generates an improved method which incorporates many desirable features of the several methods and simultaneously eliminates a number of undesirable features of the other methods.

Iterative, Intercolumnar Correlational Analysis

Iterative, Intercolumnar Correlational Analysis was developed out of a theory of types (McQuitty and Clark, 1968). A type is defined as a category of objects of such a nature that every object in the category possesses a common and unique combination of characteristics; every object in the category possesses all of these characteristics, and no object not in the category possesses all of the characteristics; a nonmember may possess some but not all of the characteristics.

Prior to the development of the iterative method many other methods for the isolation of types, as defined above, had been developed by the author. All of these methods, including the iterative method, start with a matrix of interassociations between objects. Every object is assessed in terms of selected characteristics and the relation of every object with every other object is recorded in terms of a numerical index to yield a matrix of interassociations between objects.

¹ This investigation was supported by Public Health Service Research Grant No. MH 14070-02 from the National Institute of Mental Health.

² Portions read in a Symposium on Recent Developments in Typological Analysis at the American Psychological Association, San Francisco, September, 1968.

The matrix of interassociations is analyzed in a fashion designed to classify the objects into categories which fulfill the above definition of types. The methods of analysis, prior to the iterative one, begin by classifying objects into many categories at the bottom level of classification; each category contains only a few objects and the members of each category have relatively many characteristics in common. As the analysis proceeds objects are classified into larger and larger categories at successively higher and higher levels with the members of the categories agreeing on fewer and fewer characteristics. The consolidation at each successive level is realized, for the most part, by combining categories of the next lower level. Errors made at a lower level can be carried to a higher level.

In most of the methods, the classificatory decisions are based on the highest entry in each column in the matrix of interassociations. This procedure neglects the many other entries in a matrix, which might be helpful in increasing the validity of the decisions.

In addition to the general purpose of creating an improved typological method, Iterative, Intercolumnar Correlational Analysis was directed to two specific purposes, viz., (a) to increase the validity of classification decisions, and (b) to utilize all of the entries of the matrix in making these decisions. It has, however, yet another unique characteristic. It classifies objects into a hierarchical system from top down rather than bottom up. The method divides the original matrix into two submatrices, each submatrix into two additional submatrices, and thus the process continues until at the bottom level every object is usually separated as a single object.

The Method

In order to apply the method, a matrix of interassociations between objects is required. The first step is to compute the correlation between the corresponding entries of any two columns, i and j . This index is a measure of the extent to which two objects vary jointly in their correlations with the other objects of the matrix. It is called the first intercolumnar correlation between Objects i and j . The first intercolumnar correlation is computed for every object with every other object to yield the first intercolumnar correlation matrix.

The second intercolumnar correlation between Objects i and j is obtained by computing the correlation between the corresponding

entries of Objects i and j of the first intercolumnar correlation matrix. It is computed for every object with every other object to produce the second intercolumnar correlation matrix.

As the process of generating new intercolumnar correlation matrices proceeds, a matrix is usually obtained which contains only correlations of $+1$ and -1 . There are usually two sets of $+1$'s. Each set of $+1$'s defines a submatrix. The -1 's mediate between objects of the two submatrices.

Table 1 reports a matrix of original associations to which Iterative, Intercolumnar Correlational Analysis was applied. The first, third, and fifth intercolumnar correlation matrices are reported in Tables 2, 3, and 4 respectively. Column 1 of Table 4 shows that one submatrix is composed of Object 1 and all of the other objects having a correlation of $+1$ with it. Object 2, which has a correlation of -1 with Object 1, defines the other submatrix; it is composed of Object 2 and all of the other objects which have a correlation of $+1$ with it. Each of the objects of one submatrix has a correlation of -1 with each of the objects of the other submatrix.

In continuing the analysis, the above procedures are applied to the submatrices. For this purpose, the $+1$'s of the submatrices are replaced by the corresponding entries from the original large matrix. Every submatrix is generally divided in the same fashion as just outlined for the original matrix. The steps are continued until the analysis is completed.

Some Weaknesses and Improvements

Two of the possible weaknesses of the above method are that it does use all of the data in a matrix or submatrix and it classifies from the top down. Any error made in any bifurcation is not corrected in the further analysis.

Some indices are smaller than other indices and are, therefore, less reliable. An alternative approach is to develop a method designed to use only the more reliable indices.

One approach would be to limit the computation of intercolumnar correlations to the higher indices of every matrix or submatrix being analyzed. This approach has the disadvantage of shortening the range and thereby tending to lower the reliability of the intercolumnar correlations.

TABLE 1
Agreement Scores between Objects

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		20	29	20	25	24	20	13	29	20	23	18	28	28	24	30	28	16	22	20
2	20		20	25	17	15	26	20	15	25	13	26	20	20	27	21	25	20	18	25
3	29	20		19	26	34	23	13	33	20	30	18	27	31	26	34	32	19	30	22
4	20	25	19		22	18	25	20	18	22	18	27	20	19	25	19	22	21	19	25
5	25	17	26	22		24	17	23	23	16	28	20	18	23	20	21	24	14	26	14
6	24	15	34	18	24		20	12	33	17	32	15	29	29	23	30	28	16	32	20
7	20	26	23	25	17	20		14	19	30	18	26	18	22	25	23	26	24	16	28
8	13	20	13	20	23	12	14		12	15	16	18	13	11	21	11	13	15	18	12
9	29	15	33	18	23	33	19	12		20	29	15	30	28	26	33	25	17	29	21
10	20	25	20	22	16	17	30	15	20		14	20	21	21	28	22	24	27	15	31
11	23	13	30	18	28	32	18	16	29	14		16	25	24	21	29	22	13	31	16
12	18	26	18	27	20	15	26	18	15	20	16		15	17	23	15	19	21	18	21
13	28	20	27	20	18	29	18	13	30	21	25	15		27	26	30	26	17	23	23
14	28	20	31	19	23	29	22	11	28	21	24	17	27		24	30	27	19	23	22
15	24	27	26	25	20	23	25	21	26	28	21	23	26	24		25	24	22	24	27
16	30	21	34	19	21	30	23	11	33	22	22	15	30	30	25		28	17	26	23
17	28	25	32	22	24	28	26	13	25	24	22	19	26	27	24	28		18	22	22
18	16	20	19	21	14	16	24	15	17	27	13	21	17	19	22	17	18		14	30
19	22	18	30	19	26	32	16	18	29	15	31	18	23	23	24	26	22	14		16
20	20	25	22	25	14	20	28	12	21	31	16	21	23	22	27	23	22	30		

Note.—Data for this table taken from McQuitty & Clark, 1968.

TABLE 2
First Intercorrelations of Agreement Scores of Table 1

	1	2	3	4	5	6	7	8	9	10
1	+1.00	-0.25	+0.87	-0.42	+0.40	+0.85	-0.04	-0.29	+0.85	-0.08
2	-0.25	+1.00	-0.45	+0.84	-0.48	-0.39	+0.76	+0.13	-0.31	+0.73
3	+0.87	-0.45	+1.00	-0.51	+0.51	+0.95	-0.23	-0.35	+0.94	-0.19
4	-0.42	+0.84	-0.51	+1.00	-0.62	-0.58	+0.63	+0.41	-0.56	+0.63
5	+0.40	-0.48	+0.51	-0.62	+1.00	+0.60	-0.49	-0.00	+0.52	-0.66
6	+0.85	-0.39	+0.95	-0.58	+0.60	+1.00	-0.32	-0.28	+0.95	-0.26
7	-0.04	+0.76	-0.23	+0.63	-0.49	-0.32	+1.00	+0.01	-0.21	+0.86
8	-0.29	+0.13	-0.35	+0.41	-0.00	-0.28	+0.01	+1.00	-0.33	+0.10
9	+0.85	-0.31	+0.94	-0.56	+0.52	+0.95	-0.21	-0.33	+1.00	-0.25
10	-0.08	+0.73	-0.19	+0.63	-0.66	-0.26	+0.86	-0.10	-0.25	+1.00
11	+0.74	-0.42	+0.85	-0.61	+0.73	+0.92	-0.49	-0.25	+0.87	-0.39
12	-0.47	+0.74	-0.57	+0.92	-0.47	-0.52	+0.55	+0.56	-0.54	+0.57
13	+0.87	-0.31	+0.88	-0.51	+0.47	+0.81	+0.06	-0.45	+0.90	-0.04
14	+0.93	-0.29	+0.92	-0.41	+0.35	+0.85	-0.01	-0.29	+0.92	-0.01
15	+0.23	+0.51	+0.01	+0.22	-0.36	+0.05	+0.61	-0.42	+0.07	+0.63
16	+0.90	-0.40	+0.93	-0.48	+0.46	+0.90	-0.12	-0.34	+0.94	-0.10
17	+0.84	-0.10	+0.76	-0.26	+0.22	+0.64	+0.16	-0.25	+0.74	+0.09
18	-0.19	+0.75	-0.40	+0.65	-0.61	-0.39	+0.84	-0.08	-0.33	+0.89
19	+0.69	-0.65	+0.79	-0.61	+0.77	+0.87	+0.36	-0.19	+0.81	-0.45
20	-0.02	+0.60	-0.16	+0.42	-0.50	-0.24	+0.85	+0.00	-0.14	+0.93

Note.—Data for this table taken from McQuitty and Clark, 1968.

TABLE 2 (Continued)

	11	12	13	14	15	16	17	18	19	20
1	+0.74	-0.47	+0.87	+0.93	+0.23	+0.90	+0.84	-0.19	+0.69	-0.02
2	-0.42	+0.74	-0.31	-0.29	+0.51	-0.40	-0.10	+0.75	-0.65	+0.60
3	+0.85	-0.57	+0.88	+0.92	+0.01	+0.93	+0.76	-0.40	+0.79	-0.16
4	-0.61	+0.92	-0.51	-0.41	+0.22	-0.48	-0.26	+0.65	-0.61	+0.42
5	+0.73	-0.47	+0.47	+0.35	-0.36	+0.46	+0.22	-0.61	+0.77	-0.50
6	+0.92	-0.52	+0.81	+0.85	+0.05	+0.90	+0.64	-0.39	+0.87	-0.24
7	-0.49	+0.55	+0.06	-0.01	+0.61	-0.12	+0.16	+0.84	-0.36	+0.85
8	-0.25	+0.56	-0.45	-0.29	-0.42	-0.34	-0.25	-0.08	-0.19	+0.00
9	+0.87	-0.54	+0.90	+0.92	+0.07	-0.94	+0.74	-0.33	+0.81	-0.14
10	-0.39	+0.57	-0.04	-0.01	+0.63	-0.10	+0.09	+0.89	-0.45	+0.93
11	+1.00	-0.58	+0.68	+0.74	-0.12	+0.73	+0.58	-0.46	+0.95	-0.38
12	-0.58	+1.00	-0.41	-0.41	+0.17	-0.38	-0.18	+0.53	-0.60	+0.45
13	+0.68	-0.41	+1.00	-0.91	+0.33	-0.93	+0.76	-0.13	+0.68	+0.03
14	+0.74	-0.43	+0.91	+1.00	+0.27	+0.95	+0.90	-0.19	+0.69	+0.06
15	-0.12	+0.17	+0.33	+0.27	+1.00	-0.30	+0.44	+0.61	-0.23	+0.67
16	+0.73	-0.38	+0.93	+0.95	+0.30	+1.00	-0.83	-0.17	+0.70	-0.01
17	+0.58	-0.18	+0.76	+0.90	-0.44	-0.83	+1.00	+0.02	+0.52	+0.22
18	-0.46	+0.53	-0.13	-0.19	+0.61	-0.17	+0.02	+1.00	-0.51	+0.89
19	+0.95	-0.60	+0.68	+0.69	-0.23	+0.70	+0.52	-0.51	+1.00	-0.38
20	-0.38	+0.45	+0.03	+0.06	+0.67	-0.01	+0.22	+0.89	-0.38	+1.00

TABLE 3
Third Iteration of the Agreement Scores of Table 1

	1	2	3	4	5	6	7	8	9	10
1	+1.00	-0.99	+1.00	-1.00	+0.98	+1.00	-0.97	-0.95	+1.00	-0.96
2	-0.99	+1.00	-0.99	+1.00	-1.00	-1.00	+1.00	+0.89	-0.99	+0.99
3	+1.00	-0.99	+1.00	-1.00	+0.99	+1.00	-0.98	-0.94	+1.00	-0.97
4	-1.00	+1.00	-1.00	+1.00	-0.99	-1.00	+0.99	+0.92	-1.00	+0.98
5	+0.98	-1.00	+0.99	-0.99	+1.00	+0.99	-1.00	-0.88	+0.99	-1.00
6	+1.00	-1.00	+1.00	-1.00	+0.99	+1.00	-0.98	-0.93	+1.00	-0.98
7	-0.97	+1.00	-0.98	+0.99	-1.00	-0.98	+1.00	+0.85	-0.98	+1.00
8	-0.95	+0.89	-0.94	+0.92	-0.88	-0.93	+0.85	+1.00	-0.94	+0.84
9	+1.00	-0.99	+1.00	-1.00	+0.99	+1.00	-0.98	-0.94	+1.00	-0.97
10	-0.96	+0.99	-0.97	+0.98	-1.00	-0.98	+1.00	+0.84	-0.97	+1.00
11	+0.99	-1.00	+1.00	-1.00	+1.00	+1.00	-0.99	-0.92	+1.00	-0.99
12	-1.00	+1.00	-1.00	+1.00	-0.99	-1.00	+0.98	+0.93	-1.00	+0.98
13	+1.00	-0.98	+1.00	-0.99	+0.97	+1.00	-0.96	-0.96	+1.00	-0.95
14	+1.00	-0.98	+1.00	-0.99	+0.98	+1.00	-0.96	-0.96	-0.95	-0.96
15	-0.80	+0.89	-0.83	+0.86	-0.91	-0.84	+0.93	+0.59	-0.83	+0.94
16	+1.00	-0.99	+1.00	-1.00	+0.98	+1.00	-0.97	-0.95	+1.00	-0.96
17	+1.00	-0.97	+0.99	-0.98	+0.96	+0.99	-0.94	-0.98	+0.99	-0.93
18	-0.97	+1.00	-0.98	+0.99	-1.00	-0.98	+1.00	+0.85	-0.98	+1.00
19	+0.99	-1.00	+1.00	-1.00	+1.00	+1.00	-0.99	-0.91	+1.00	-0.99
20	-0.95	+0.99	-0.97	+0.98	-0.99	-0.97	+1.00	+0.82	-0.97	+1.00

Note.—Data for this table taken from McQuitty and Clark, 1968.

TABLE 3 (Continued)

	11	12	13	14	15	16	17	18	19	20
1	+0.99	-1.00	+1.00	+1.00	-0.80	+1.00	+1.00	-0.97	+0.99	-0.95
2	-1.00	+1.00	-0.98	-0.98	+0.89	-0.99	-0.97	+1.00	-1.00	+0.99
3	+1.00	-1.00	+1.00	+1.00	-0.83	+1.00	+0.99	-0.98	+1.00	-0.97
4	-1.00	+1.00	-0.99	-0.99	+0.86	-1.00	-0.98	+0.99	-1.00	+0.98
5	+1.00	-0.99	+0.97	+0.98	-0.91	+0.98	+0.96	-1.00	+1.00	-0.99
6	+1.00	-1.00	+1.00	+1.00	-0.84	+1.00	+0.99	+0.98	+1.00	-0.97
7	-0.99	+0.98	-0.96	-0.96	+0.93	-0.97	-0.94	+1.00	-0.99	+1.00
8	-0.92	+0.93	-0.96	-0.96	+0.89	-0.95	-0.98	+0.85	-0.91	+0.82
9	+1.00	-1.00	+1.00	+1.00	-0.83	+1.00	-0.99	-0.98	+1.00	-0.97
10	-0.99	+0.98	-0.95	-0.96	+0.94	-0.96	-0.93	+1.00	-0.99	+1.00
11	+1.00	-1.00	+0.99	+0.99	-0.86	+0.99	+0.98	-0.99	+1.00	-0.98
12	-1.00	+1.00	-1.00	-1.00	+0.84	-1.00	-0.99	+0.98	-1.00	+0.97
13	+0.99	-1.00	+1.00	+1.00	-0.79	+1.00	+1.00	-0.96	+0.99	-0.94
14	+0.99	-1.00	+1.00	+1.00	-0.80	+1.00	-1.00	-0.97	-0.87	+0.95
15	-0.86	+0.84	-0.79	-0.80	+1.00	-0.80	-0.75	+0.93	+0.99	-0.95
16	+0.99	-1.00	+1.00	+1.00	-0.80	+1.00	+1.00	-0.97	+0.99	-0.95
17	+0.98	-0.99	+1.00	+1.00	-0.75	+1.00	+1.00	-0.94	+0.98	-0.92
18	-0.99	+0.98	-0.96	-0.97	+0.93	-0.97	-0.94	+1.00	-0.99	+1.00
19	+1.00	-1.00	+0.99	+0.99	-0.87	+0.99	+0.98	-0.99	+1.00	-0.98
20	-0.98	+0.97	-0.94	-0.95	+0.95	-0.95	-0.92	+1.00	-0.98	+1.00

TABLE 4
Fifth and Final Iteration from Table 1

	1	2	3	4	5	6	7	8	9	10
1	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0
2	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	+1.0
3	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0
4	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	+1.0
5	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0
6	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0
7	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	+1.0
8	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	+1.0
9	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0
10	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	+1.0
11	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0
12	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	+1.0
13	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0
14	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0
15	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	+1.0
16	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0
17	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0
18	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	+1.0
19	+1.0	-1.0	+1.0	-1.0	+1.0	+1.0	-1.0	-1.0	+1.0	-1.0
20	-1.0	+1.0	-1.0	+1.0	-1.0	-1.0	+1.0	+1.0	-1.0	+1.0
	11	12	13	14	15	16	17	18	19	20
1	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	-1.0
2	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0
3	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	-1.0
4	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0
5	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	-1.0
6	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	-1.0
7	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0
8	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0
9	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	-1.0
10	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0
11	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	-1.0
12	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0
13	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	-1.0
14	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	-1.0
15	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0
16	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	-1.0
17	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	-1.0
18	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0
19	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	+1.0	-1.0	+1.0	-1.0
20	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	-1.0	+1.0	-1.0	+1.0

Note.—Data for this table taken from McQuitty and Clark, 1968.

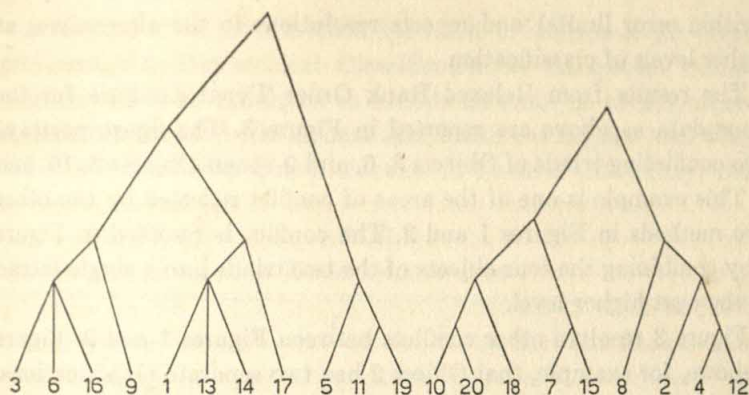


Figure 2. A Hierarchical Classification of the Objects by Multiple Linkage Analysis. (Reprinted from McQuitty, 1969.)

16. The new method, on the other hand, as can be seen in Figure 2, first classifies Objects 3, 6, and 16 (rather than 9) into a triad which is then joined by Object 9.

The differences in classification are based on small numerical differences in the data in relation to unique features of the methods of analysis; many of the differences in results of this kind are within chance errors.

A Revision of Rank Order Typal Analysis

A method is needed which portrays options of the above kind in classifying objects. It should show the indices on which the options are based, and it should resolve all conflicts at higher levels of classification.

Another new method by the author has merit for this purpose. It is a revision of Rank Order Typal Analysis and is called Relaxed Rank Order Typal Analysis (McQuitty 1971). Both the original and the revised methods derive from a definition of types. A type is a category of objects of such a nature that every object of the category is more like every other object of the category than it is like any object of any other category. This strict definition of a type precludes the finding of any but a few small types in most sets of empirical, psychological data. The revised method relaxes the definition sufficiently to isolate types in most sets of data.

An added feature of relaxing the definition is that the method gives alternative classifications for certain objects which are highly similar

(within error limits) and reports resolutions to the alternatives at higher levels of classification.

The results from Relaxed Rank Order Typal Analysis for the same data as above are reported in Figure 3. The figure portrays two conflicting triads of Objects 3, 6, and 9 versus Objects 3, 16, and 9. This example is one of the areas of conflict reflected by the other two methods in Figures 1 and 2. The conflict is resolved in Figure 3 by combining the four objects of the two triads into a single tetrad at the next higher level.

Figure 3 resolves other conflicts between Figures 1 and 2. Figure 3 shows, for example, that Object 2 has two separate classifications, one corresponding with its classification in Figure 1 and the other corresponding with its classification in Figure 2. Figure 3 shows the level at which these two alternative classifications are resolved.

*A Method Which Adjusts the Classification Criterion
to the Requirements of the Data*

Figure 3 does not, however, resolve all of the conflicts between Figures 1 and 2. A method is needed whereby it is easy to adjust the criterion for classification to the requirements of the data.

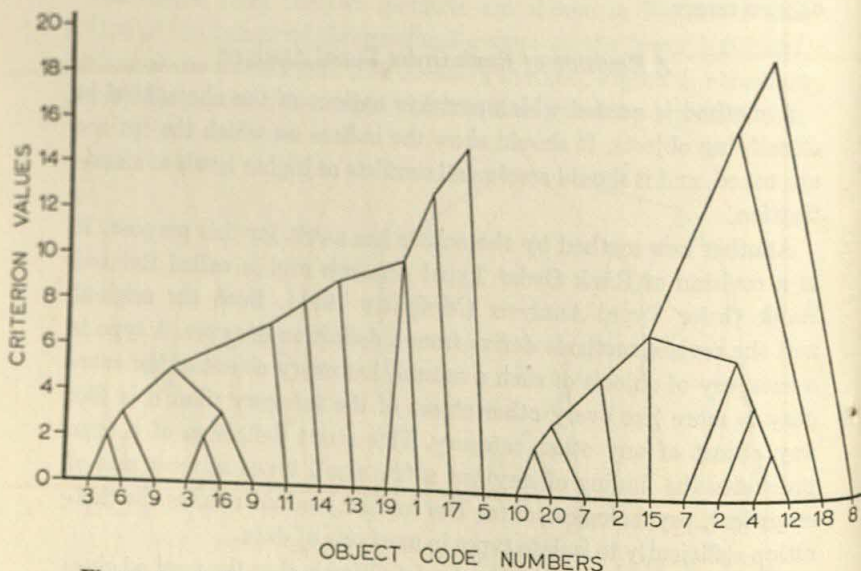


Figure 3. A Hierarchical Classification by Relaxed Rank Order Typal Analysis.

A revision of an older method has merit in relation to the present problems, viz., Hierarchical Classification by Reciprocal Pairs. A revision of it was developed to include an analysis of ties, where i is highest with j and j is highest with i , and i is highest also with k and k is highest with i , as illustrated in Table 5 (McQuitty, 1966;

TABLE 5
A Tie in Reciprocal Pairs

	i	j	k
i	—	—	—
j	—	—	—
k	—	—	—

— Highest entry in a column.

McQuitty, Price, and Clark, 1967). The problem of ties is solved by classifying i with j and i with k .

That this method works effectively with data containing ties is illustrated in Figure 4, which reports the results from applying it

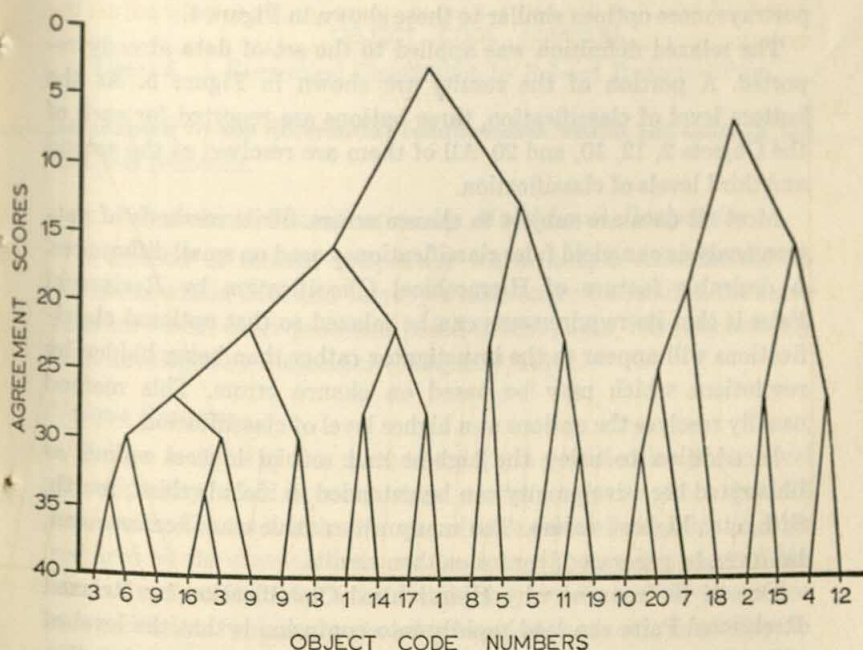


Figure 4. A Hierarchical Classification by Reciprocal Pairs for Data with Ties.

to the same data as was analyzed by the other methods. Because of ties, Object 3 is classified separately with Objects 6 and 16; Object 9 is classified separately with Categories 3-6 and 3-16, and at the next higher level of classification, i.e., the third level, all of these conflicts are resolved because Objects 3, 6, 9, and 16 enter a common tetrad. In addition to its classification with Categories 3-6 and 3-16, Object 9 is also classified separately with Object 13. This latter conflict is resolved at the fourth level of classification. This method resolves problems, as just illustrated, in a fashion similar to Relaxed Rank Order Typal Analysis.

An advantage of the reciprocal pairs method is that it portrays clearly all of the indices on which the classification is based, including the tied values, and reveals to the investigator the multiple classifications and usually the ways in which they are resolved; the objects in conflict usually enter a common category, sooner or later.

The definition of a reciprocal pair can be relaxed so that a pair can be accepted as reciprocal if i is either highest or second highest with j and j is either highest or second highest with i . This approach portrays more options similar to those shown in Figure 4.

The relaxed definition was applied to the set of data already reported. A portion of the results are shown in Figure 5. At the bottom level of classification, three options are reported for each of the Objects 2, 12, 10, and 20. All of them are resolved at the second and third levels of classification.

Most all data are subject to chance errors. Strict methods of pattern analysis can yield false classifications based on small differences. A desirable feature of Hierarchical Classification by Reciprocal Pairs is that its requirements can be relaxed so that optional classifications will appear to the investigator rather than being hidden by resolutions which may be based on chance errors. This method usually resolves the options at a higher level of classification.

In addition to using the highest and second highest values as illustrated here, reciprocity can be extended to include third, fourth, fifth, etc., highest values. Too many alternative classifications can, however, be generated; confusion then results.

One of the reasons why Hierarchical Classification by Relaxed Reciprocal Pairs can lead rapidly into confusion is that the level of relaxation is spread throughout the data. If appropriate relaxation is to be realized and confusion avoided, the level of relaxation must

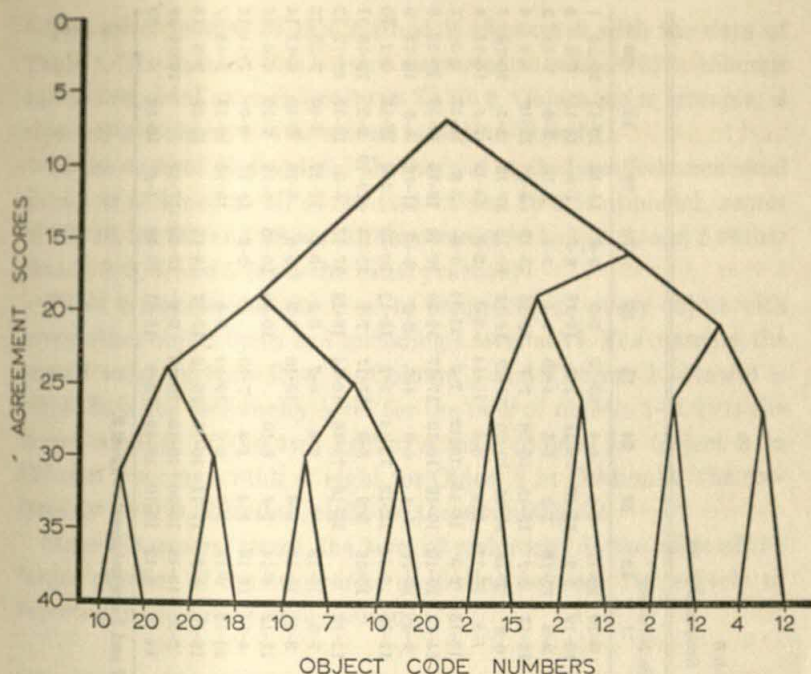


Figure 5. A Hierarchical Classification by Relaxed Reciprocal Pairs.

be adapted to the differential requirements within the data as the analysis proceeds.

Hierarchical Classification by Multi-Level Reciprocity

A method of relaxed reciprocity which adapts differentially to the needs within data and thereby avoids confusion, and at the same time solves the other problems posed in this paper, has just recently been developed by the author (McQuitty, 1970).

Unique Features

Unique features of this method are (1) a specification of the level of reciprocity used at every stage of the analysis, (2) a gradual increase in the level of reciprocity (more relaxed classification) as required by the characteristics of the data, (3) an adjustment of the level of reciprocity to both the validity of the data and the size of the categories into which objects are classified, and (4) an ability to reject an object for classification because it is inappropriate to the categories generated by the other objects.

TABLE 6
Multiple Level Reciprocal Pairs

	Object Code Numbers																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2	9-12		8-8	10-12	4-7	9-9	12-12	11-19	5-5	11-12	9-10	10-17	4-4	4-4	10-10	3-3	2-4	14-18	10-11	14-14
3	12-12		14-14	2-4	15-16	17-17	3-3	3-9	17-17	5-5	18-19	2-2	13-13	15-15	2-2	14-14	8-8	7-9	13-15	5-5
4	2-8*	9-14		13-16	2-10	1-1	9-12	11-19	1-3	11-14	3-6	10-18	5-9	1-5	4-10	1-1	1-4	8-16	3-6	9-13
5	12-12	4-4	16-16		10-10	14-17	6-6	8-10	15-17	7-7	12-17	1-1	13-13	16-16	7-7	16-16	13-13	5-9	13-13	5-5
6	7-7	16-16	10-10	6-10		9-9	17-17	1-7	11-11	16-17	6-6	7-12	15-15	10-10	19-19	14-14	10-10	17-18	5-5	18-18
7	8-9	17-17	1-1	17-17	2-9		12-12	15-19	1-8	15-15	1-3	16-17	3-6	3-6	14-14	5-5	2-8	14-16	1-5	14-14
8	12-12	2-3	12-12	2-6	15-17	12-12		10-19	14-14	2-2	12-15	2-3	15-15	12-12	7-7	11-11	6-6	3-8	16-18	3-3
9	19-19	9-9	19-19	10-10	7-7	19-19	19-19		19-19	17-17	14-14	10-10	19-19	19-19	17-17	19-19	16-16	13-13	19-19	8
10	2-5	17-17	3-3	17-17	7-11	2-2	14-14	15-19		11-13	4-5	16-17	1-4	4-8	4-9	2-2	8-10	11-16	4-5	12-12
11	10-10	19-19	6-6	17-17	1-6	3-3	15-15	7-14	5-5	19-19		17-19	7-11	14-14	1-3	13-13	10-10	2-4	18-18	1-1
12	17-17	2-8	18-18	1-1	12-12	17-17	3-3	5-10	17-17	11-11	14-15		15-15	9-9	8-8	17-17	7-7	13-13	19-19	2-2
13	4-4	9-13	9-9	10-13	14-15	6-6	15-15	11-19	4-4	9-12	7-9	16-18	18-18	18-18	14-14	18-18	17-17	5-5	13-13	12-12
14	4-4	9-15	5-5	13-16	7-10	6-6	11-12	18-19	8-8	9-14	8-8	14-18	5-6		4-7	3-3	6-7	11-17	8-10	7-10
15	8-10	1-2	10-10	2-7	12-19	11-14	6-7	2-17	9-9	3-3	11-17	4-14	7-7	8-10	10-10	10-10	4-16	7-10	4-4	15
16	1-8	8-14	1-1	13-16	11-14	5-5	9-11	18-19	1-8	7-13	4-7	16-18	1-3	5-5	7-10		2-8	11-17	5-9	7-11
17	4-4	4-8	4-4	16-13	5-10	8-8	3-6	11-19	10-10	6-10	10-13	9-17	7-7	6-6	10-10	8-8		10-18	10-13	9-13
18	18-18	9-9	16-16	9-9	18-18	16-16	8-8	8-16	16-16	4-4	18-19	5-5	17-17	16-16	16-16	17-17	18-18		19-19	2-2
19	11-11	15-15	6-6	13-13	2-5	3-3	18-18	5-13	5-5	17-18	2-2	10-13	10-10	10-10	10-10	9-9	13-13	17-19		16-16
20	12-14	4-5	13-13	2-6	18-18	12-14	2-3	15-19	12-12	1-1	14-16	5-12	10-10	12-12	2-4	11-11	13-13	1-8	16-16	

* 2-8, for example, means Object 3 ranks second with Object 1 and Objects 1 and 3 are reciprocal at Level 8.

See Row 1—Column 3.

NOTE.—Reprinted from McQuitty, 1970.

The novel feature of this method is illustrated with the data of Table 1. The data of Table 1 are converted to ranks within columns and are reported as rank orders in Table 6. Object 3, for example, is reported in Column 1 to be second most like Object 1.

In the case of tied values, the highest rank (smallest numerical value) is assigned to all of the tied values. In this approach, scores of 30, 29, 29, 29, and 28 would have ranks of 1, 2, 2, 2, and 5 rather than 1, 3, 3, 3, and 5 (as is the usual practice).

Table 6 reports also the level of reciprocity of every object with every other object up to and including a level of 19. For example, the second value in both Row 1—Column 3 and Column 1—Row 3 is eight. It is the reciprocity level for the pair of objects 1-3. It is the larger number of the two ranks, a rank of two for Object 3 in Column 1 versus a rank of eight for Object 1 in Column 3. The reciprocity level is, therefore, eight for these two objects.

Stated in general terms, the level of reciprocity is the value of the larger number of the two ranks mediating between two objects as reported in their respective columns.

Two Versions

The *reciprocity level* is used as the *classification criterion*. It can be applied in two slightly different ways: (1) successive linkages, or (2) core attachments.

Successive linkages. In *successive linkages*, one starts with a criterion of one. Any two objects which have a reciprocity of one between them constitute a beginning category. Every other object which has a one with either of the two members of the category is added to the category.

Other pairs of objects with a criterion of one between them are sought and built up in the same fashion as above, using a criterion of one.

After all such pairs have been exhausted and each has been built to its maximum under a criterion of one, the criterion is increased by units of one and each category built to its maximum under each size of the criterion. The analysis is complete when every object has been assigned in terms of its smallest reciprocity value.

Core attachments. In *core attachments*, the cores are defined as categories generated at any level of the classification criterion; the

most basic approach is to restrict them to the largest categories which can be realized by a criterion of one.

The core approach differs from *successive linkages* by the fact that any object brought into the core must satisfy the criterion with respect to every object in the core, not just some one of them. If the initial pair is composed of Objects i and j and we are using a criterion of one, then Object x qualifies to join the pair if and only if it has a reciprocity level of one with each i and j . Likewise, object y then qualifies if and only if it has a reciprocity of one with each i , j , and x . By contrast, *successive linkages* requires entry objects to satisfy the criterion with at least one object of the category.

After core categories have been isolated and built to their maximum sizes, all other objects are assigned to them using the method of *successive linkages*.

In either approach, multiple classification due to near ties can be introduced by relaxing the classification criterion.

Results

Results with the data of Table 1 are shown in Figure 6 for the successive linkage version and in Figure 7 for the core assignment version. With the present data they give identical results in terms of the categories into which the objects are classified.

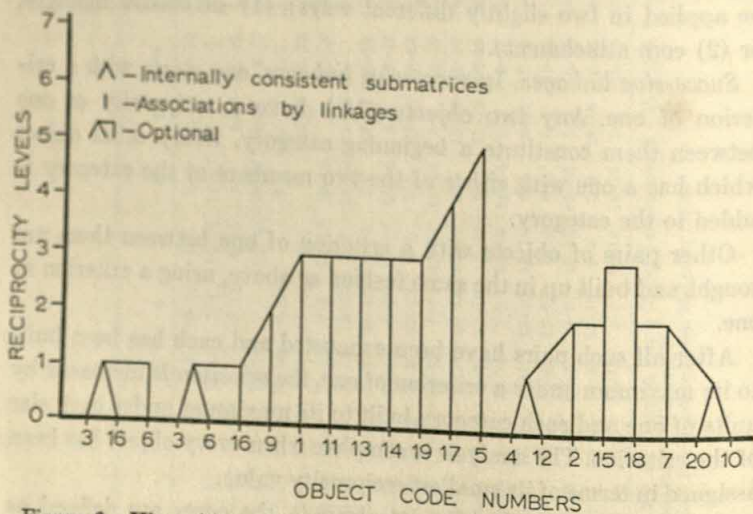


Figure 6. Hierarchical Classification by Successive Linkages. (Reprinted from McQuitty, 1970.)

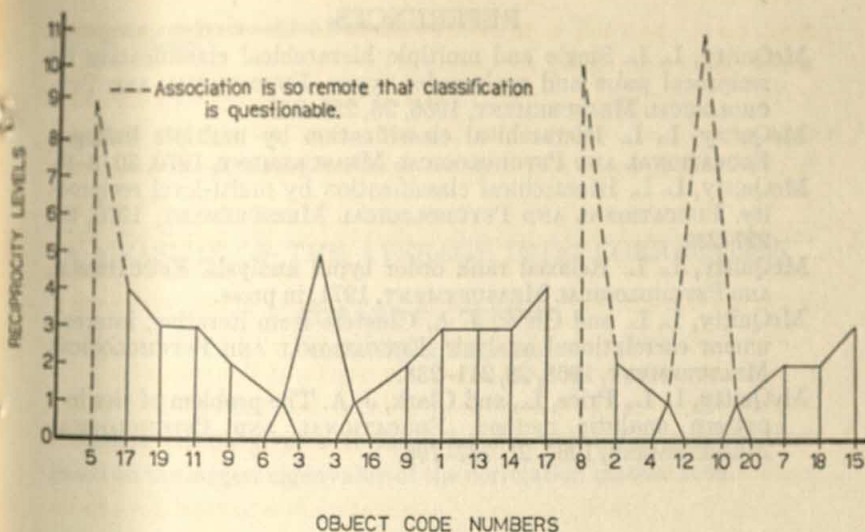


Figure 7. Hierarchical Classification by Core Assignments. (Reprinted from McQuitty, 1970.)

Desirable Features

Hierarchical Classification by Multi-Level Reciprocity, as just outlined, has many desirable features, including: (1) setting the classification criteria at successively higher and higher levels as required by the data, (2) classifying from bottom up (a small initial criterion) or top down (a large initial criterion), (3) yielding the same results irrespective of the starting point, (4) analyzing fairly large matrices by pencil and paper, (5) portraying all classification decisions and the objective basis for them, (6) reporting the internal consistency of the results, (7) excluding "misfits," and (8) assigning objects either to central cores exclusively or to central cores and their extensions as illustrated here.

Summary

The application of selected methods of pattern analysis to a particularly difficult set of data illustrates strengths and weaknesses of the methods and serves as a basis for the development of a new method which combines several desirable features and eliminates certain undesirable features.

REFERENCES

- McQuitty, L. L. Single and multiple hierarchical classification by reciprocal pairs and rank order types. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1966, 26, 253-265.
- McQuitty, L. L. Hierarchical classification by multiple linkages. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1970, 30, 3-19.
- McQuitty, L. L. Hierarchical classification by multi-level reciprocity. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1970, 30, 227-239.
- McQuitty, L. L. Relaxed rank order typal analysis. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1971, in press.
- McQuitty, L. L. and Clark, J. A. Clusters from iterative, intercolumnar correlational analysis. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 211-238.
- McQuitty, L. L., Price, L., and Clark, J. A. The problem of ties in a pattern analytic method. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 787-796.

A MEASURE OF THE AVERAGE INTERCORRELATION

EDWARD E. CURETON

University of Tennessee

KAISER (1968) gives a formula for the average intercorrelation based on the largest eigenvalue of the correlation matrix. It is

$$\hat{r} = \frac{E - 1}{p - 1},$$

where E is the largest eigenvalue and p is the number of variables. It occurred to the writer that a simple approximation to this formula might be based on the first centroid of the correlation matrix.

Kaiser considers a correlation matrix $R(p \text{ by } p)$, with unities on the diagonal and equal values r elsewhere. Solving for r in terms of p and the largest eigenvalue of R , he obtains the formula given above.

The column sums of R will each be $1 + (p - 1)r$, and the total sum will be p times this value or $p[1 + (p - 1)r]$. Corresponding to the largest eigenvalue, the sum of squares of the first centroid factor loadings will be

$$\sum f^2 = \frac{p[1 + (p - 1)r]^2}{p[1 + (p - 1)r]} = 1 + (p - 1)r,$$

from which

$$r = \frac{\sum f^2 - 1}{p - 1},$$

which is precisely Kaiser's formula with E replaced by $\sum f^2$.

Using Hotelling's classical example,

$$R = \begin{bmatrix} 1.000 & .698 & .264 & .081 \\ .698 & 1.000 & -.061 & .092 \\ .264 & -.061 & 1.000 & .594 \\ .081 & .092 & .594 & 1.000 \end{bmatrix}$$

$$\begin{array}{cccccc} 2.043 & 1.729 & 1.797 & 1.767 & 7.336 \end{array}$$

we find that

$$\sum f^2 = \frac{\sum_i (\sum r)^2}{\sum_i \sum_i r} = \frac{13.515}{7.336} = 1.842,$$

from which $r = .281$. Kaiser finds for this same problem that $r = .282$, and the still simpler arithmetic mean of the off-diagonal entries is .278.

REFERENCE

Kaiser, H. F. A measure of the average intercorrelation. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 245-247.

SELF-CLAIMED AND TESTED KNOWLEDGE¹

RALPH F. BERDIE

Student Life Studies
University of Minnesota

THE traditional method for observing whether or not a person knows something is to develop a test or examination which provides an opportunity for him to demonstrate his knowledge. The person is asked to indicate, through recall or recognition, the answer to a question and the person asking the question then decides regarding the correctness or appropriateness of the answer. The decision regarding the person's knowledge is a function of what he actually knows, the way the question is asked, and the way the judgment is made concerning the correctness of the answer. The assumption is made that if a question is asked of a person and he answers the question in the proper way, he has the information, and if he does not answer the question properly, he does not have the information.

An alternative method to determine whether or not a person knows something is to ask him if he knows it. Thus one can present an individual with a list of laws, principles, persons, or facts and ask the person to check or otherwise indicate the ones with which he is familiar and the degree of his familiarity.

Many people do not like to take tests and become frightened or anxious when faced by a test. Tests also can demand much time to develop, administer, and score. On the other hand, few people are reluctant to tell another that they do or do not know something and the collection of such information can be done fairly quickly and economically.

The motivation of the respondent and his perception of the rea-

¹ The author acknowledges the assistance of Mr. Gary R. Hanson in developing the test, collecting the data, and analyzing the results.

son for which he is being questioned most likely effect his responses, regardless of their mode. Revealing the extent of ones information to a professor offering pellets of high grades may be quite different from responding for a graduate student collecting dissertation data.

The author, in an attempt to observe the reported experiences of students, both before and after entering college, developed a set of experience inventories and asked students to describe their experiences (private music lessons, attendance at lectures and concerts, membership in youth organizations, and so on). As part of this experience inventory, three separate lists of relatively well-known persons were devised, a list of authors, a list of painters, and a list of other public figures, including businessmen, politicians, entertainers, and athletes. Students were asked to indicate that they had never heard of the person, that they had heard of the person but had no other experience regarding him, or that they had read a book by the person, seen a picture painted by the person, or knew who the person was. The responses of students in different colleges were different, and the responses of students before and after two years of college were different. Thus the method apparently reflected differential experiences.

The question still remained, however, as to the correspondence that existed between what students said they knew and what they actually knew, as shown by more traditional achievement examination.

Method

From the experience inventory a list of 12 persons well-known in public life, 13 well-known authors, and 14 well-known painters was prepared. The names included were: Bishop Pike, Henry Miller, John Gardner, Melina Mercouri, Van Cliburn, Thomas Watson, James Conant, Shirley Booth, Werner Von Braun, Francis Spellman, Stewart Udall, Lorne Green, Albert Camus, Fyodor Dostoyevski, Ayn Rand, Henry James, James Baldwin, D. H. Lawrence, Leo Tolstoi, William Golding, J. D. Salinger, Jean Paul Sartre, James Joyce, Francois-Marie Voltaire, James Michener, Edgar-Hilaire Degas, Henri de Toulouse-Lautrec, Peter Paul Rubens, Grant Wood, Velasquez, El Greco, Botticelli, Manet, Andy Warhol, Cezanne, Salvador Dali, Jackson Pollack, Vincent Van Gogh, and Raphael.

For the well-known men, students were asked to respond in one of

three ways: know who he is, have heard of him but cannot identify him, have never heard of him. The responses for the authors consisted of: read a book by him, heard of him but have not read a book by him, have never heard of him. The responses for painters consisted of: seen a picture by him, heard of him but have not seen a picture by him, have never heard of him. Students were instructed to place check marks in the appropriate positions and the purpose of the research presented before they completed the checklist was described as to determine how much students knew of these things.

The achievement examination consisted of 40 items and the stem of each item included the names of the well-known persons. Five alternatives were presented for each item. Examples are: Bishop Pike is known for his experiences with (a) LSD (b) hypnosis (c) group meditation (d) speaking with the dead (e) anxiety perception; Which of these novels was written by Albert Camus? (a) *Pere Goriot* (b) *Walden Two* (c) *One Flew Over the Cuckoo's Nest* (d) *The Great Gatsby* (e) *The Stranger*; Salvador Dali painted (a) "The Last Supper" (b) "Young Beggar" (c) "The Steamship" (d) "View of Toledo" (e) "Titus."

Total scores were obtained for the checklist and for the test, and three additional subscores were obtained for each instrument, one based on artists, one based on authors, and one based on public figures.

Two samples were studied. One contained 84 males and 80 females, mostly sophomores, who were drawn from the subject pool of the second semester of general psychology in the spring quarter of 1969. Each of these subjects received two "Course grade points" for participating in the research. The second sample consisted of 17 males and 35 females who lived in a coeducational freshman dormitory and who volunteered to take these tests in order to receive one dollar for their effort.

The instructions for each instrument were printed on the first page. The experience checklist was distributed and subjects were instructed to read the sheet of instructions and begin immediately. No time limit was set and most subjects finished in 7 to 10 minutes. After the checklist was completed and before the test was given to the students, they were told that the research was designed to compare the two methods of observing what students know and that after they had completed the test the experimenters would com-

pare their responses to the checklist to the answers they provided on the test. Subjects then were told to read the printed test instructions and begin immediately. Most subjects required between 15 and 25 minutes to complete the test.

Within both samples, analyses were completed separately for men and women. The experience checklist was scored by assigning a weight of three to the category of knowing who the person was, a weight of two to the category of having heard of the person, and a weight of one to the category of never having heard of the person. The total score for the checklist consisted of the sum of all of these items. The score for the achievement test consisted of the number of items answered correctly.

The number of subjects who checked each category on the checklist was determined, the number of subjects who checked each category on the checklist and also provided the correct answer for the corresponding item on the achievement test was observed, and the per cent of individuals who checked each category and also checked the correct answer was determined. The product-moment correlation coefficients were computed between the total scores and for the three subtest scores for each sample.

Students' responses to the test items cannot provide an absolute indication of their knowledge or lack of knowledge about the person. Subjects were asked to identify the name of a person with one of several possible facts about that person and some of the subjects might have known much about a person but not known the fact presented. For example, the alternatives presented for the novelist, James Michener, included the names of four books that Michener did not write and the title of a book he did write, *Tales of the South Pacific*. A student might not have known that Michener wrote that particular book, but he may have read another book by him and have known quite a lot about him. Thus, the test gives only one of several possible indications as to the students' knowledge regarding these people.

Results

Table 1 presents the intercorrelations between the test scores and the checklist scores.

The correlations between total scores on the test and the checklist ranged from .47 to .74. The correlations for the three largest samples

TABLE 1

Intercorrelations between Scores on the Test and Scores on the Checklist for the Four Samples

	<i>N</i>	Artists (14 items)	Authors (13 items)	Public Figures (12 items)	Total (39 items)
CLA males	84	.07	.44	.76	.74
CLA females	80	-.05	.69	.67	.67
Dormitory males	17	-.07	.40	.72	.47
Dormitory females	35	-.03	.30	.40	.65

were .65 and above. These three correlations for the total scores are statistically significant beyond the .01 level, the correlation for the smallest group is significant beyond the .05. The correspondence between the two measures is more than faintly observable.

The subtest correlations are greatest for the public figures and nonexistent for the artists. This may be due in part to the relatively greater experience students have with names of public figures and the little experience they have in the field of fine arts.

The results on Table 1 suggest that a checklist provides a rough but acceptable means for determining how much students, as a group, know about some things but not about others.

The next analysis was of relationships between responses to individual items on the checklist and corresponding items on the test. For example, of the 80 women in the psychology pool sample, 43 said they knew who Bishop Pike was, 28 said they had heard of him but could not identify him further, and 9 said they had never heard of him. Of the 43 who said they knew who he was, 65 per cent answered correctly the test item regarding Pike. Of the 28 who said they had heard of him, 43 per cent answered the test item correctly. Of the 9 who said they had never heard of him, 11 per cent answered the item correctly. Chance alone would provide that 20 per cent would answer the item correctly insofar as there were five alternative answers. For each of the 39 items, the percentage of students who said they knew the items and who answered the test item correctly was determined. Then of the students who indicated they had heard the item but had no further experience with it, the percentage who answered the test item was observed. Finally, of the students who

indicated they did not know the item, the percentage who answered correctly the test item was determined.

This analysis first was done for the 12 items regarding public figures. For the 80 women, the percentages indicating on the checklist that they knew who these people were ranged from 10 to 76, with Lorne Green being the best known person. The percentages indicating they had heard of these people but knew no more about them ranged from 1 to 28. The percentages indicated they had never heard of these people ranged from 2 to 53. Conant was the least known person. For this group of women, and considering only the items pertaining to public figures, the median percentage correct on the test was 56. Of the students who checked they knew who these people were, the median percentage correct on the test was 85. For the people who checked that they were acquainted with the names of these people but could not identify them, the median percentage correct on the test was 53. For the people who said they did not know who they were, the median percentage correct was 18.

For the authors' item, the median percentage correct on the test was 66. Considering only those persons who said they had read books by these authors, the median percentage correct on the test was 74. For those who indicated they knew the persons but had not read books by them, the median percentage correct on the test was 68. For the subjects who said they did not know who these authors were, the median percentage correct was 20.

For the artists, the median percentage correct for the total group was 39. For the subjects who said they had seen pictures by these artists, the median percentage correct was 52. For the subjects who recognized the names but had not seen pictures by these persons, the median percentage correct was 23, and for the subjects who said they did not recognize the names of the artists, the median percentage correct was 22.

Thus, for the subjects who described their knowledge as nil, the median percentage correct on the tests was chance. For the most part, people who said they knew more about the person also performed better on corresponding items.

An inference regarding the validity of the method can be obtained by examining results from the original experience inventory. Among the list of names of thirty authors were included the names of four persons who were, as far as this writer knows, nonexistent

and among the listed names of forty artists were included the names of four persons who were not known as artists. Among a group of entering freshman men, from 0 to 3 per cent reported they had read books by the nonexistent authors and from 5 to 25 per cent indicated they had heard of these persons. From 71 to 93 per cent reported they had never heard of these persons; the remainder of the responses were unclassifiable. From 0 to 8 per cent of the students indicated they had seen pictures by the nonexistent painters and from 4 to 19 per cent reported they had heard of these. From 72 to 96 per cent reported they had never heard of these nonexistent painters. Insofar as the students were under no external pressure to lie, the responses indicating experiences with the nonexistent persons might be attributed to erroneous associations with the names of existing persons. For example, one of the nonexistent author names was Samuel Green. Students might have associated this with someone like Samuel Butler or Graham Green. One of the names presented was Gilbert Deck, a name that apparently does not resemble that of any well-known author, for only 1 per cent of the students reported they had read a book by him and only an additional 5 per cent indicated they had actually heard of him. Similarly, the name of Paul Fondley, presented as an artist, was recognized by only 4 per cent of the students and none of the students claimed they had seen a picture by him. These sketchy results suggest that not much purposeful distortion was reflected by the experience inventory.

Conclusion

The results suggest that for survey purposes, asking people whether or not they possess information may provide a satisfactory means for observing whether or not they know something. The effectiveness of the checklist method may be quite dependent on the content and also on the level of familiarity the subjects have with the content.

The results also suggest that if a person on a checklist indicates he does not know something, his lack of information will be verified by an achievement test. The checklist method may not be quite as adequate in showing whether or not people who think they know something actually do know it. In a sense, we have here an excellent method for determining the extent of a person's ignorance, perhaps

a less satisfactory method for determining the extent of his knowledge.

This conclusion has practical significance. Surveys of amount of knowledge and information possessed by members of a group can be simplified by successive screening. Large numbers of persons can be asked to indicate on a checklist what they know and then achievement tests can be given only to those claiming knowledge. The assumption is that those who say they do not know really are ignorant and further testing of them is unnecessary. How acceptable this assumption is depends in part on the motivation of respondents.

A SIGNIFICANCE TEST FOR BISERIAL r^1

EDWARD ALF² AND NORMAN ABRAHAMS

Naval Personnel and Training Research Laboratory
San Diego, California 92152

THE biserial correlation coefficient, r_b , is a statistic often used in test construction and validation. A disadvantage of r_b is that its exact sampling distribution in small samples is not known. Thus significance testing becomes a problem; and approximate methods suitable only for large samples or limited ranges of difficulty level or criterion dichotomization (p) and magnitude of r_b must be used (see, e.g. Guilford, 1965, p. 319; Walker and Lev, 1953, p. 271). An advantage of the point-biserial correlation coefficient, r_{pb} , is that its exact sampling distribution is known.

The present paper will show that under the null hypothesis $\rho = 0$, the assumptions of r_b and r_{pb} coincide. This makes it possible to derive exact, small sample significance tests for r_b from the corresponding known distribution of r_{pb} .

The point-biserial correlation coefficient is used to determine the relation between a dichotomous variable, X , and a continuous variable, Y . The development of r_{pb} assumes the continuous variable, Y , is normally distributed in each X category, and the variance of Y is the same in each X category. No assumption is made as to the distribution of the dichotomous variable, but generalization is made only to a universe of samples of size N having the same fixed number of cases Np and Nq in the dichotomous categories (Walker and Lev, 1953, p. 271).

¹ The opinions expressed are those of the authors and do not necessarily reflect those of the Navy Department.

² Also at San Diego State College.

For example, if we compute r_{pb} on random samples of size 10 from some given population, where $Np = 6$ cases fall in one X category and $Nq = 4$ cases fall in the other X category, we would expect r_{pb} ($df = 8$) to exceed $\pm .632$ one time in twenty when $\rho = 0$. The value of .632 can be found in a table of significance values for r_{pb} (see, e.g. Guilford, 1965, p. 580).

The biserial correlation assumes we are sampling from a bivariate normal population; and the two X categories are formed by a dichotomization of the X continuum. It is often true that the distribution of Y will not be the same in each X category under the assumptions of r_b (Walker and Lev, 1953, p. 271). However, when $\rho = 0$, the distribution of Y will be independent of X . Therefore, the distribution of Y will be identical in each X category; and if the (X, Y) distribution is bivariate normal, then Y will of necessity be normally distributed in each X category, and will have the same variance in each X category. Thus, when $\rho = 0$, the assumptions of r_b and r_{pb} are the same.

For any given values of N , p and q , r_b will be a constant times r_{pb} ; that is:

$$r_b = (r_{pb}) \frac{\sqrt{pq}}{y},$$

where y is the ordinate of the normal curve at the point of dichotomization. Therefore, it is possible to determine significant values for r_b directly from the significant values of the corresponding r_{pb} .

In the above example, for instance, where $p = .6$, we have:

$$r_b = (r_{pb}) \frac{\sqrt{(.6)(.4)}}{.3863} = r_{pb}(1.268).$$

Thus, if $r_{pb} = \pm .632$ is significant at the five per cent level, then $r_b = \pm (.632)(1.268) = \pm .801$ will also be significant at the five per cent level.

Using the relation between r_{pb} and r_b , it is possible to determine significant values for r_b at any desired p value and for any desired sample size. Table 1 presents two-tailed significance values for r_b at the five and one per cent levels of significance for selected sample sizes ranging from 7 to 1,000, and for p -values ranging from .05 to .95 at five per cent intervals. This table should be useful to the investigator who wishes to assess the significance of his obtained biserial correlations.

TABLE 1

The 5 (Roman Type) and 1 (Italics) Per cent Significance Levels for r_s

Sample Size	Point of Dichotomization									
	.50 (.50)	.55 (.45)	.60 (.40)	.65 (.35)	.70 (.30)	.75 (.25)	.80 (.20)	.85 (.15)	.90 (.10)	.95 (.05)
7	.946	.948	.957	.972	.994	—	—	—	—	—
8	.886	.888	.896	.910	.931	.963	—	—	—	—
9	.835	.838	.845	.858	.878	.908	.952	—	—	—
10	.792	.794	.801	.814	.833	.861	.903	.968	—	—
11	.755	.757	.763	.775	.794	.820	.860	.922	—	—
12	.722	.724	.730	.742	.759	.785	.823	.882	.985	—
13	.693	.695	.701	.712	.729	.753	.790	.847	.945	—
14	.667	.669	.675	.686	.702	.725	.761	.815	.910	—
15	.644	.646	.652	.662	.677	.700	.734	.787	.879	—
16	.623	.625	.631	.640	.655	.678	.711	.762	.850	—
17	.604	.606	.611	.621	.635	.657	.689	.738	.824	—
18	.587	.589	.594	.603	.617	.638	.669	.717	.800	.990
19	.571	.573	.578	.587	.600	.621	.651	.698	.779	.963
20	.556	.558	.563	.571	.585	.605	.634	.680	.759	.938
21	.543	.544	.549	.557	.570	.590	.618	.663	.740	.915
22	.530	.531	.536	.544	.557	.576	.604	.647	.723	.893
23	.518	.519	.524	.532	.545	.563	.590	.633	.706	.873
24	.507	.508	.513	.521	.533	.551	.578	.619	.691	.855
25	.496	.498	.502	.510	.522	.540	.566	.607	.677	.837
26	.487	.488	.492	.500	.512	.529	.555	.595	.664	.820
27	.477	.479	.483	.490	.502	.519	.544	.583	.651	.805
28	.469	.470	.474	.481	.493	.509	.534	.573	.639	.790
29	.460	.462	.466	.473	.484	.500	.525	.562	.628	.776
	.590	.591	.596	.606	.620	.641	.672	.720	.804	.994

TABLE 1 (Continued)

Sample Size	Point of Dichotomization									
	.50 (.50)	.55 (.45)	.60 (.40)	.65 (.35)	.70 (.30)	.75 (.25)	.80 (.20)	.85 (.15)	.90 (.10)	.95 (.05)
30	.452	.454	.458	.465	.476	.492	.516	.553	.617	.700
	.580	.582	.587	.596	.610	.631	.661	.709	.791	.900
40	.391	.392	.396	.402	.411	.425	.446	.478	.533	.630
	.505	.506	.511	.518	.531	.549	.575	.617	.688	.800
60	.319	.320	.322	.327	.335	.346	.363	.389	.435	.530
	.414	.415	.418	.425	.435	.450	.471	.505	.564	.680
100	.246	.247	.249	.253	.259	.268	.281	.301	.336	.410
	.321	.322	.325	.330	.338	.349	.366	.393	.438	.540
125	.220	.221	.223	.226	.232	.239	.251	.269	.300	.370
	.288	.288	.291	.296	.302	.313	.328	.351	.392	.480
150	.201	.202	.203	.206	.211	.218	.229	.246	.274	.330
	.263	.263	.266	.270	.276	.285	.299	.321	.358	.440
200	.174	.174	.176	.179	.183	.189	.198	.213	.237	.290
	.228	.228	.230	.234	.239	.247	.259	.278	.310	.380
250	.156	.156	.157	.160	.164	.169	.177	.190	.212	.260
	.204	.204	.206	.209	.214	.221	.232	.249	.278	.340
300	.142	.142	.144	.146	.149	.154	.162	.173	.194	.230
	.186	.186	.188	.191	.195	.202	.212	.227	.254	.310
350	.131	.132	.133	.135	.138	.143	.150	.161	.179	.220
	.172	.173	.174	.177	.181	.187	.196	.210	.235	.290
400	.123	.123	.124	.126	.129	.134	.140	.150	.168	.200
	.161	.161	.163	.165	.169	.175	.183	.197	.219	.270
500	.110	.110	.111	.113	.116	.120	.125	.134	.150	.180
	.144	.144	.146	.148	.151	.156	.164	.176	.196	.240
1000	.078	.078	.079	.080	.082	.084	.089	.095	.106	.130
	.102	.102	.103	.105	.107	.111	.116	.124	.139	.170

Note.—A dash (—) indicates r_s would have to exceed 1.00 to be significant.

REFERENCES

- Guilford, J. P. *Fundamental statistics in psychology and education*. (4th ed.). New York: McGraw-Hill, 1965.
- Walker, H. M. and Lev, J. *Statistical inference*. New York: Henry Holt & Co., 1953.

STATISTICAL CONTROL OF "IMPURITY" IN THE ESTIMATION OF TEST RELIABILITY

K. H. LU

Department of Biostatistics
University of Oregon Dental School
Portland, Oregon

TRADITIONALLY, the reliability of a test is defined as the ratio of the variance of true test scores to the variance of the observed test scores. Numerous articles have been written on the concept of reliability and the ways of estimation. The three typical techniques of estimation are (1) the Kuder-Richardson formula 20 (Kuder and Richardson, 1937), (2) the Hoyt analysis of variance method (Hoyt, 1941), and (3) the Rulon split half method (Rulon, 1939). It has been shown that the Kuder-Richardson formula 20 and Hoyt's method are algebraic equivalents. The reliabilities of many tests have been computed by these methods since their introduction some 30 years ago.

Unfortunately, from the theoretical point of view, each of these methods suffers from a certain amount of impurity. Consequently, these methods arrive at the correct estimates only when the impurities are accidentally absent from the data. For instance, Hoyt and Krishnaiah (1960) investigated an analysis of variance model where the item-subject interaction was assumed absent. Under various assumptions regarding the nature of the effects of the items and the subjects, they found some impurity exists in this simple model.

It is the purpose of this paper to:

1. render precise definitions of the concepts of reliability for the single item, the whole test, and the relationship between them in the least-squares sense;
2. elucidate the sources of impurities present in the current methods of estimation;

3. suggest an estimating procedure such that the resultant reliabilities are free from such impurities; and
4. give the statistical definition of significant reliability which serves as the necessary condition for a "meaningful" reliability.

The Definitions of Item and Test Reliabilities

Suppose that the observed score of test item j by a subject S_i is given by

$$y_{ij} = \mu + S_i + e_{ij} \quad \begin{matrix} i = 1, 2, \dots, m \\ j = 1, 2, \dots, n \end{matrix} \quad (1)$$

where μ = the mean of true scores

S_i = the true deviation from μ for the i th subject, such that

$$\sum_{i=1}^m S_i = 0$$

e_{ij} = an observational random error which is normally independently distributed $(0, \sigma^2)$.

We see that the true score of the i th subject is $\mu + S_i$. We shall define the reliability per item as the intraclass correlation.

$$r_I = \frac{\sigma_S^2}{\sigma_S^2 + \sigma^2} \quad (2)$$

Now let us consider the problem of the definition of reliability for a test of n items.

Let

$$y_{i.} = \frac{1}{n} \sum_{j=1}^n y_{ij} = \frac{1}{n} \left[n\mu + nS_i + \sum_{j=1}^n e_{ij} \right];$$

we have

$$y_{i.} = \mu + S_i + e_{i.} \quad (3)$$

Suppose that μ is known or effectively estimable; we wish to obtain scores of the form $r_i(y_{i.} - \mu)$ as the best estimate of the true scores of the subject S_i in the least-squares sense. This requires the choice of r_i such that

$$\Phi = E[r_i(y_{i.} - \mu) - S_i]^2$$

is a minimum.

Expand Φ , we have

$$\begin{aligned}\Phi &= r_i^2 E(y_{i.} - \mu)^2 - 2r_i E(y_{i.} - \mu)(S_i) + E(S_i)^2 \\ &= r_i^2 \left(\sigma_s^2 + \frac{\sigma^2}{n} \right) - 2r_i \sigma_s^2 + \sigma_s^2\end{aligned}$$

Differentiate Φ with respect to r_i and set the derivative equal to zero, we have

$$\frac{d\Phi}{dr_i} = 2r_i \left(\sigma_s^2 + \frac{\sigma^2}{n} \right) - 2\sigma_s^2 = 0$$

and then

$$r_i = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma^2}{n}} = \frac{n\sigma_s^2}{n\sigma_s^2 + \sigma^2}. \quad (4)$$

We shall call the r_i thus derived the *reliability of the test*.

It is of interest to note that r_i is the resultant reliability of applying the Spearman-Brown formula to the reliability per item, r_I . It can be shown as follows:

From equation (2), we have

$$r_I = \frac{\sigma_s^2}{\sigma_s^2 + \sigma^2} = \frac{\sigma_s^2/\sigma^2}{\sigma_s^2/\sigma^2 + 1}$$

Therefore,

$$\frac{\sigma_s^2}{\sigma^2} = \frac{r_I}{1 - r_I}$$

From (4), we have

$$\begin{aligned}r_i &= \frac{n\sigma_s^2}{n\sigma_s^2 + \sigma^2} = \frac{\sigma_s^2/\sigma^2}{\sigma_s^2/\sigma^2 + 1/n} \\ &= \frac{r_I/(1 - r_I)}{r_I/(1 - r_I) + 1/n} = \frac{nr_I}{1 + (n - 1)r_I}\end{aligned} \quad (5)$$

which is the familiar Spearman-Brown formula (Spearman 1904).

The test reliability as defined by equation (4) is totally consistent with the definition given by previous investigators,

$$r_i = \frac{\text{Var (true test score)}}{\text{Var (observed test score)}}.$$

The Impurity in the Current Estimation Methods

Unfortunately, the computation procedures used in computing r_k , such as the Kuder-Richardson 20, or its equivalents, the Hoyt analysis of variance method and the Rulon split-half method, are all inappropriate to some extent, so that they seldom truly approach

$$\frac{nr_s^2}{nr_s^2 + \sigma^2},$$

as they should. In the narrative to follow, we shall discuss the inappropriateness of these computation procedures.

Let us consider a test of n items be given to m subjects and the results summarized as two-factor factorial table, where the subjects and the items have m and n levels respectively.

TABLE 1

*Results of an n -item Test Taken by m Subjects * x_{ij} = score of j th item by i th subject*

Subjects	Items				Σ
	1	2..... j n			
1	x_{11}	x_{12}	x_{1j}	x_{1n}	$x_{1.}$
2	x_{21}	x_{22}	x_{2j}	x_{2n}	$x_{2.}$
.
i	x_{i1}	x_{i2}	x_{ij}	x_{in}	$x_{i.}$
.
m	x_{m1}	x_{m2}	x_{mj}	x_{mn}	$x_{m.}$
Σ	$x_{.1}$	$x_{.2}$	$x_{.j}$	$x_{.n}$	$x_{..}$

The Kuder-Richardson 20 and the Hoyt method may be illustrated by the following mathematical model in analysis of variance:

$$x_{ij} = \mu + S_i + I_j + e_{ij} \quad (6)$$

where

- x_{ij} = the score of the i th subject on the j th item
- = 1 if answered correctly
- = 0 otherwise.

μ = an effect common to all x_{ij} 's

S_i = a component specifically associated with the i th subject

I_j = an effect specifically associated with the j th item

e_{ij} = a random error independently normally distributed with mean zero and variance σ^2 .

In order to discuss the issues intelligently, we must first discuss the designs of analysis of variance appropriate for our purposes.

1. *The fixed effect model*

In this model, it is understood that only the particular item included in the test and the group of subjects are of special interest to the examiner. The reliability of the test results is only applicable to those specific items and subjects, no statistical inferences on the populations at large are to be made. Unless the purpose of the test is so stated as above, the usage of the fixed effect model should be avoided.

2. *The random effect model*

This model assumes that (1) the items of the test are a random sample of size n from a population of items; (2) the subjects are a random sample of size m from a population of subjects. The results are to be inferred to the populations of items and subjects at large. While it is reasonable to assume the m subjects are a random sample, seldom if ever is a test written which would select test items strictly at random. For most tests, the items are specifically selected for the expressed purpose of measuring the subjects' "level of achievement" of the area. Furthermore, if the same test is to be used over and over again with various groups of subjects, it is then anything but a random sample. The random effect model's assumption that the items are a random sample, therefore, exerts very strict restriction on the usefulness of the random effect model. However, on occasions where the items can indeed be considered as a random sample, the random effect model of course can be used.

3. *The mixed-effect model with Item-Subject interaction used as experimental error*

In view of the foregoing discussion, a mixed-effect model would appear more suitable for our purpose. By a mixed-effect model, it is meant that: (1) the items are considered fixed, and (2) the subjects are a random sample from a population of subjects. However, difficulties in the correct estimation of reliability remains. This is due to the fact that the test results are considered as a two-factor (item and subject) factorial without direct estimate of the experimental error σ^2 . In the analysis to follow, we shall show the consequences in each of these three models, resulting in impurities in the estimation of reliability.

TABLE 2

The Components of Variance for the Fixed, the Random and the Mixed Models

Sources of Variation	d.f.	M.S.	Fixed Model	M.S. is Estimate of Random Model	Mixed Model
Items	$n - 1$	M_4	$\sigma^2 + m\sigma_I^2$	$\sigma^2 + \sigma_{IS}^2 + m\sigma_I^2$	$\sigma^2 + \sigma_{IS}^2 + m\sigma_I^2$
Subjects	$m - 1$	M_3	$\sigma^2 + n\sigma_S^2$	$\sigma^2 + \sigma_{IS}^2 + n\sigma_S^2$	$\sigma^2 + n\sigma_S^2$
$I \times S$	$(n - 1)(m - 1)$	M_2	$\sigma^2 + \sigma_{IS}^2$	$\sigma^2 + \sigma_{IS}^2$	$\sigma^2 + \sigma_{IS}^2$
Error	0	M_1	σ^2	σ^2	σ^2
Total	$nm - 1$				

The partitions of mean squares of the analysis of variance according to the three models are given in Table 2. We also see that the error mean square is not estimable because it has zero degrees of freedom. According to Hoyt's formula, the reliability is calculated as

$$r_t = \frac{M_3 - M_2}{M_3}$$

Thus for the fixed and mixed models,

$$r_t = \frac{n\sigma_S^2 - \sigma_{IS}^2}{n\sigma_S^2 + \sigma^2},$$

and for the random model,

$$r_t = \frac{n\sigma_S^2}{n\sigma_S^2 + \sigma_{IS}^2 + \sigma^2}.$$

Unless $\sigma_{IS}^2 = 0$, the Kuder-Richardson formula 20 or Hoyt's methods will always under-estimate r_t .

Rulon's formula is equivalent to the following analysis of variance table involving only the subject's scores of the halves. Note that

TABLE 3

The Components of Variance for Rulon's Split-Halves Model

Sources of Variation	d.f.	M.S.	Mean Square is Estimate of:
Halves	1	M_4	$\sigma^2 + \frac{n}{2}\sigma_{HS}^2 + \frac{nm}{2}\sigma_H^2$
Subjects	$m - 1$	M_3	$\sigma^2 + n\sigma_S^2$
$(H \cdot S)$	$m - 1$	M_2	$\sigma^2 + \frac{n}{2}\sigma_{HS}^2$
Total	$2m - 1$		

there are $(m \times n)$ observations, hence $mn - 1$ degrees of freedom for analysis. But the Rulon's method only utilized $2m - 1$ of them.

From Table 3, we have

$$r_1 = \frac{M_3 - M_2}{M_3}$$

$$= \frac{n\sigma_s^2 - \frac{n}{2}\sigma_{HS}^2}{\sigma^2 + n\sigma_s^2}$$

Again it suffers from a theoretical "impurity" σ_{HS}^2 which tends to estimate the reliability incorrectly.

In the event that $\sigma_{HS}^2 = 0$, then M_2 is in fact an estimate of σ^2 , but it is estimated by $(m - 1)$ degrees of freedom, sometimes may result in overestimating or underestimating reliability; whereas from the available data, a more efficient estimation of error mean square by $m(n/2 - 1)$ degrees of freedom is available but not utilized. This inefficient use of available data should be avoided.

From the above analysis, it is abundantly clear the mixed effect model would serve our purpose if it can be made to provide an estimate of the experimental error σ^2 . In the narrative to follow we shall show such a model.

The Appropriate Design and Model for the Estimation of Reliability

In view of the foregoing analysis, it becomes apparent that in order to obtain the proper estimate of reliability, two salient points must be taken into consideration for the construction of a mathematical model: (1) the inclusion of the item-subject interaction term and (2) The direct estimation of σ^2 and subsequent estimation of σ_s^2 . Guttman (1945) demonstrated that in order to estimate the reliability coefficient, at least two tests are needed, thus leading to the procedure of the test-retest case. In our present case, we seek a method of estimation where only one test is required. In order to satisfy these requirements, one must in some way partition the test into at least two "comparable" parts in order to provide an independent estimate of σ^2 . There are various ways of partitioning the test into two parts, for example it may be done by item contents or item difficulty. While it is difficult to pair items by "comparable" contents, it is a simple matter to pair them by difficulty from the

results of the test, since the error variance σ^2 is a function of item difficulty (Lord, 1957). In fact, if the partition is done on the difficulty basis, it results in a reduction of the size of error variance σ^2 (Osborn, 1969). Let a sample of m subjects be tested by a test of n different items, (we should require that n be an even number), in a manner such that if the item is answered correctly, the subject receives one point, and if answered incorrectly, the subject receives nothing. We shall arrange the n items in a descending order according to their respective numbers of correct answers, the item with the largest number of correct answers is listed first, the item with the second largest number of correct answers next, and the item with the least number of correct answers last. The list thus obtained is then a list of the n items by degree of difficulty in an ascending order.

Since the degree of difficulty is monotonic in nature as appeared in the list, we may systematically assign all the odd-numbered items to the first half and all the even-numbered items to the second half of the test. Thus, we have $n/2$ pairs of items with one member of each pair in each half of the test. It appears reasonable to view the halves thus obtained as being comparable in the sense that each item in one half has a counterpart in the other with comparable degree of difficulty. (The systematic assignment will give the odd half perhaps a greater mean than the even half, but would not affect the variances of the halves.) The test results can be considered as being stratified in pairs according to item difficulty. As we shall see later, this stratification enables us to obtain an estimate of reliability without the entanglement present in the current methods.

Let the model of the test be defined as follows:

$$y_{ijk} = \mu + H_k + S_i + (SH)_{ik} + I_j + (IS)_{ij} + e_{ijk}$$

where $i = 1, 2, \dots, m$; $j = 1, 2, \dots, n/2$; $k = 1, 2$, and

y_{ijk} = the score of the i th subject on the j th item in the k th half
 = 1 if answered correctly
 = 0 otherwise

μ = a common effect shared by all y_{ijk} 's

H_k = a specific effect associated with the k th half; the halves are fixed effects such as a result of systematic designation such that $\sum H_k = 0$.

S_i = a component specifically associated with the i th subject.
The S_i 's are assumed normally independently distributed with mean zero and variance σ_s^2 .

I_i = a specific effect associated with the j th item pair. The I_i 's are purposefully chosen items, therefore, considered fixed and $\sum I_i = 0$.

HS_{ik} = a component associated with the specific combination of the i th subject and the k th half such that $\sum_{k=1}^m HS_{ik} = 0$
Note that we do not require $\sum_{k=1}^2 HS_{ik} = 0$

SI_{ij} = a component associated with the specific combination of the i th subject and the j th item pair, such that $\sum_{i=1}^{n/2} SI_{ij} = 0$; again we do not require $\sum_{i=1}^m SI_{ij} = 0$

e_{ijk} = a random error normally and independently distributed with mean zero and variance σ^2 .

The analysis of variance and the components of the mean squares are listed in Table 4. The test of significance of reliability is the

$$H_0: \sigma_s^2 = 0 \text{ by}$$

$$F = M_s/M_1 \text{ with } (m-1) \text{ and } m(n/2-1) \text{ degrees of freedom.}$$

The computation of reliability for the test is

TABLE 4
The Components of Variance for Mixed Model with Items Stratified According to Difficulty

Sources of Variation	d.f.	M.S.	Mean Square is Estimate of:
Halves	1	M_6	$\sigma^2 + \frac{n}{2}\sigma_{HS}^2 + \frac{mn}{2}\sigma_H^2$
Subjects	$m-1$	M_5	$\sigma^2 + n\sigma_s^2$
HS	$m-1$	M_4	$\sigma^2 + \frac{n}{2}\sigma_{HS}^2$
Item-pairs	$\frac{n}{2}-1$	M_3	$\sigma^2 + 2\sigma_{IS}^2 + 2n\sigma_I^2$
IS	$(m-1)(\frac{n}{2}-1)$	M_2	$\sigma^2 + 2\sigma_{IS}^2$
Error	$m(\frac{n}{2}-1)$	M_1	σ^2
Total	$mn-1$		

$$r_t = \frac{n\sigma_s^2}{n\sigma_s^2 + \sigma^2} = \frac{M_s - M_1}{M_s},$$

and for the reliability per item is

$$r_I = \frac{\sigma_s^2}{\sigma_s^2 + \sigma^2} = \frac{M_s - M_1}{M_s + M_1(n-1)}.$$

Significant Reliability vs Meaningful Reliability

By the definitions of reliability,

$$r_t = \frac{n\sigma_s^2}{n\sigma_s^2 + \sigma^2} \quad \text{and} \quad r_I = \frac{\sigma_s^2}{\sigma_s^2 + \sigma^2}$$

we see that unless we are reasonably certain that $\sigma_s^2 > 0$, the reliability must be deemed as essentially zero. In order to assert that $\sigma_s^2 > 0$, we must rely upon the rejection of the hypothesis that $\sigma_s^2 = 0$. The F test

$$F = \frac{M_s}{M_1} = \frac{\sigma^2 + n\sigma_s^2}{\sigma^2}$$

with $(m-1)$ and $m(n/2-1)$ degrees of freedom would serve this purpose. We shall at the point define the term, significant reliability (r_t or r_I) as a reliability estimate based on data analysis where the F test concerning the $H_0: \sigma_s^2 = 0$ has been significant. A non-significant F value would suggest the need of a new test. For a reliability estimate to be meaningful, the requirement of being significant would serve as a necessary condition, though just how high the reliability has to be in order to be meaningful depends very much on the purpose of the test.

A Numerical Example

In Table 5 are the item and subjects scores of a 12-item test by 10 subjects from Guilford's book *Psychometric Methods* (Guilford, 1954).

The above example is chosen for its wide accessibility since the Guilford book is a standard reference for workers in this area.

(1) Using Kuder-Richardson formula 20, we find

$$\sum pq = 2.03, \quad \sigma_t^2 = 9.45, \quad n = 12$$

$$r_t = \left(\frac{n}{n-1} \right) \left(\frac{\sigma_t^2 - \sum pq}{\sigma_t^2} \right)$$

$$= \left(\frac{12}{11}\right)\left(\frac{9.45 - 2.03}{9.45}\right) = .857$$

(2) Using Hoyt's analysis of variance method, we have

Sources	d.f.	S.S.	M.S.
Subjects	9	7.875	.875
Items	11	9.492	.863
Error	99	12.425	.126
Total	119	29.792	

$$r_i = \frac{.875 - .126}{.875} = .857.$$

(3) Using Rulon's split-half method:

The split of the test items are 1, 3, 5, 7, 9, and 11 as the first half, and 2, 4, 6, 8, 10, and 12 as the second half as was done by Guilford; the computation is shown in Table 6.

From Table 6, we have $\sigma_d^2 = 1.05$, $\sigma_t^2 = 9.45$. Thus, according to Rulon's formula,

TABLE 6

Scores of Odd and Even Items by 10 Subjects Arranged for Rulon's Method of Estimating Reliability

X_o	X_e	$(X_o - X_e)$ d	d^2	$(X_o + X_e)$ X_t	X_t^2
2	0	+2	4	2	4
2	2	0	0	4	16
2	2	0	0	4	16
2	3	-1	1	5	25
3	2	+1	1	5	25
4	2	+2	4	6	36
4	3	+1	1	7	49
4	5	-1	1	9	81
6	5	+1	1	11	121
6	6	0	0	12	144
$\Sigma 35$	30	+5	13	65	517
145	120		Σd^2	ΣX_t	ΣX_t^2
ΣX_o^2	ΣX_e^2				
$M 3.5$	3.0	+0.5		6.5	
$\sigma^2 2.25$	3.00	1.05		9.45	
$r_{oe} = .809$					

$$r_i = 1 - \frac{\sigma_d^2}{\sigma_i^2} = 1 - \frac{1.05}{9.45} = .889.$$

The analysis of variance is presented in Table 7 We first test the

TABLE 7

Analysis of Variance and Components of Variance of the Mixed Model with Stratification According to Item Difficulty

Sources of Variation	d.f.	S.S.	M.S.	Mean Square is Estimate of
Halves	1	.2077	(M_6) .2077	
Subjects	9	7.8750	(M_5) .8750	$\sigma^2 + 12\sigma_S^2$
$H \cdot S$	9	.8750	(M_4) .0972	$\sigma^2 + 6\sigma_{HS}^2$
I	5	9.1417	(M_3) 1.8283	
$I \cdot S$	45	6.2750	(M_2) .1344	
Error	50	5.4173	(M_1) .1083	σ^2
Total	119	29.7917		

$$H_0: \sigma_{HS}^2 = 0$$

$$F = \frac{.0972}{.1083} = .8975$$

We accept $H_0: \sigma_{HS}^2 = 0$. Note that if we compute

$$r_i = \frac{M_4 - M_3}{M_3} = \frac{.8750 - .0972}{.8750} = .8889,$$

which is identical to Rulon's formula. In the present case, based on 9 degrees of freedom and the acceptance of $\sigma_{HS}^2 = 0$, the estimate of σ^2 is .0972. However, the most efficient estimate of σ^2 is the error mean square $\sigma^2 = .1083$ based on 50 degrees of freedom.

Thus we estimate

$$r_i = \frac{M_4 - M_1}{M_4} = \frac{.8750 - .1083}{.8750} = .8762,$$

and also

$$\sigma_s^2 = \frac{.8750 - .1083}{12} = .0639.$$

Thus reliability per item is

$$r_I = \frac{.0639}{.0639 + .1083} = .3711.$$

We now list the results for comparison purposes:

Methods	r_t
Kuder-Richardson 20	.8570
Hoyt's	.8570
Rulon's	.8889
Mixed-effect model	.8762

From the above demonstration, we see the underestimate of r_t by the Kuder-Richardson 20 and the Hoyt method. The overestimation of r_t by Rulon's method is due to the small degrees of freedom in the estimation of σ^2 . The mixed-effect model estimate of r_t thus calculated is free from the impurities.

One of the applications of r_t aside from its meaning as a measure of reliability is its use in estimating individual subject's true scores.

$$\begin{aligned} y_i &= \mu + S_i \\ &= \mu + r_t(\mu_i - \mu) \end{aligned}$$

Therefore,

$$\begin{aligned} \hat{y}_i &= \bar{y} + r_t(\bar{y}_i - \bar{y}) \\ &= \bar{y}(1 - r_t) + r_t\bar{y}_i \end{aligned}$$

For example, $\bar{y} = 6.50$, $r_t = .8762$, the estimated true scores of the ten subjects are: 2.56, 4.31, 4.31, 5.19, 5.19, 6.06, 6.94, 8.69, 10.44 and 11.32.

REFERENCES

- Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1964.
- Gunther, W. C. *Analysis of variance*. Englewood Cliffs: Prentice-Hall, 1964.
- Guttman, L. A. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, 10, 255-282.
- Hicks, C. R. *Fundamental concepts in the design of experiments*. New York: Holt, Rinehart and Winston, 1965.
- Hoyt, C. J. Test reliability estimated by analysis of variance. *Psychometrika*, 1939, 6, 153-160.
- Hoyt, C. J. and Krishnaiah, P. R. Estimation of test reliability by analysis of variance. *Journal of Experimental Education*, 1960, 28, 257-259.
- Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, 2, 151-160.
- Lord, F. Do tests of the same length have the same standard error

- of measurement, *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1957, 17, 510-521.
- Osborn, H. G. The effect of item stratification on errors of measurement. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 295-301.
- Ostle, B. *Statistics in research: Basic concepts and technique for research workers*. Ames: Iowa State University Press, 1958.
- Rulon, R. J. A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, 1939, 9, 99-103.
- Spearman, C. Correlation calculated from faulty data. *American Journal of Psychology*, 1904, 15, 271-295.

IS THERE AN OPTIMAL NUMBER OF ALTERNATIVES FOR LIKERT SCALE ITEMS? STUDY I: RELIABILITY AND VALIDITY

MICHAEL S. MATELL¹ AND JACOB JACOBY
Purdue University

GIVEN that rating scales are so widely used in the social sciences, both as research tools and in practical applications, determination of the optimal number of rating categories becomes an important consideration in the construction of such scales. As Garner (1960) pointed out, the basic question is whether for any given rating instrument there is an optimum number of rating categories, or at least a number of rating categories beyond which there is no further improvement in discrimination of the rated items. Garner and Hake (1951), Guilford (1954), and Komorita and Graham (1965) indicated that if we use too few rating categories, our scale is obviously a coarse one, and we lose much of the discriminative powers of which the raters are capable. Conversely, we could also grade a scale so finely that it is beyond the rater's limited powers of discrimination.

Ghiselli (1948) and Guilford (1954) contended that the optimal number of steps is a matter for empirical determination in any situation, and suggested that there is a wide range of variation in refinement around the optimal point in which reliability changes very little. Guilford felt that it may be advisable in some favorable situations to use up to 25 scale deviations. Ghiselli suggested that either reliability of measurement or ease of rating be used as a basis for the empirical determination of the optimal number of steps. Factors which affect the optimal number of rating categories, or at

¹ Now at the Procter and Gamble Company, Cincinnati, Ohio.

least the number beyond which there will be no further improvement in discrimination, are, according to Garner (1960), clearly a function of the amount of discriminability inherent in the items being rated. He suggested that there can be no single number of categories appropriate for all rating situations.

Champney and Marshall (1939) reported that under favorable rating conditions the practice of limiting rating scales to five or seven points may often give inexcusably inaccurate results. They suggested that the optimal number of steps is a function of the conditions of measurement. They also considered that, unless it could be shown that for a particular task either accuracy is not desirable or discrimination beyond seven points cannot be attained, it may be appropriate to use 18- to 24-step rating scales. Both Champney and Marshall and Guilford suggested that when the rater is trained and interested, the optimal number of steps may be in the 20-point range. The literature, however, contains few descriptions of scales employing such large numbers of rating categories.

Jahoda, Deutsch, and Cook (1951) and Ferguson (1941) opined that the reliability of a scale increases, within limits, as the number of possible alternative responses is increased. Cronbach (1950) suggested that there is no merit to increasing the reliability of an instrument unless its validity is also increased at least proportionately. He concluded that "it is an open question whether a finer scale of judgment gives either a more valid ranking of subjects according to belief, or scores more saturated with valid variance (p. 22)." Earlier, Symonds (1924), in contrast to Cronbach, contended that the problem of determining the number of steps to utilize is primarily one of reliability. He implied that optimal reliability is obtained with a 7-point scale. If more than seven steps are utilized, increases in reliability would be so small that it would not pay for the extra effort involved. However, if the raters are untrained or relatively disinterested, maximal reliability will be reached with fewer steps. Champney and Marshall (1939) suggested that nine steps in a rating scale produce the maximal reliability for trained raters. Contrary to the above suggestions, results of empirical investigations by Bendig (1954) and Komorita (1963) indicated that reliability is independent of the number of scale points employed. Komorita concluded that utilization of a dichotomous scale would not significantly decrease the reliability

of the information obtained when compared to that obtained from a multi-step scale.

Whether an increase in the number of scale points is associated with an increase in reliability, and how many scale points should be employed beyond which there would be on further meaningful increase in reliability, are both empirical questions. Studies addressed to these reliability questions have typically employed a measure of internal consistency (either split-half stepped up by the Spearman-Brown Prophecy Formula or Kuder-Richardson Formula 20). Utilization of a stability (test-retest) measure appears to be nonexistent. It should be apparent that both reliability coefficients—internal consistency *and* stability—must be assessed if meaningful and complete answers to the questions posed are to be provided.

Moreover, studies dealing with the number of alternatives problem emphasize reliability as the major, and in some instances, only criterion in the choice of the number of scale points. However, according to both Cronbach (1950) and Komorita and Graham (1965), the ultimate criterion is the effect a change in the number of scale points has on the validity of the scale. An intensive literature search failed to reveal any empirical investigation addressed to this question.

Multi-step Likert-type rating scales provide two components of information—the direction and the intensity of an individual's attitudinal composition. Peabody (1962) concluded that the total scores obtained with any Likert-type scale represent primarily the directional component, and only to a minor degree, the intensity component. Both Peabody (1962) and Cronbach (1950) suggested that differences in the intensity component primarily represent differences in response set tendencies, i.e., tendencies for subjects to use a particular degree of agreement or disagreement toward any attitudinal object regardless of the direction. Cronbach concluded that any increase in test reliability due to response set, in the final analysis, dilutes the test results and lowers its validity.

This investigation was undertaken to answer a fundamental and deceptively simple question: is there an optimal number of alternatives to use in the construction of a Likert-type scale? Of specific concern was whether variations in the number of scale alternatives affected either reliability or validity.

Method

Subjects

Four-hundred and ten undergraduate psychology students enrolled in a large midwestern university participated in this experiment. The procedure first involved selecting adjective statements for each scale point ($n = 40$), then determining the inter-rater reliability on those statements selected ($n = 10$), and, lastly, conducting the experiment proper ($n = 360$) in which 20 subjects were assigned to each of the 18 different Likert scale formats. Different samples of students attending classes in general introductory psychology, introductory applied psychology, industrial psychology, and consumer psychology were used for each segment of the study.

Scale Construction and Instruments

Anchoring verbal statements to Likert scale points has usually been conducted on an intuitive basis. In the present investigation the statements were determined empirically. Forty subjects already familiar with the technique of paired comparisons were presented 17 different statements, ranging from "I am uncertain" to "I infinitely agree," in paired comparison format, and asked to select the statement from each pair "which indicated greater agreement." A total of 136 comparisons were made by each subject. (There is no reason to believe that the results would have been any different had the instructions specified disagreement rather than agreement.) The information derived from this procedure served as the basis for selecting those statements used to construct the 18 (i.e., 2- to 19-point) Likert-type rating formats. Criteria for the selection of a statement were: (a) that it have a minimal number of reversals (less than five out of a possible 40), and (b) that it be approximately equidistant (where possible) from the statement preceding and following it. Ten of the original 17 statements came closest to meeting those criteria. These 10 statements were then presented to a new group of 10 students who were instructed to rank them in the order of increasing disagreement. (The purpose of the disagreement instructions with these subjects, in contrast to the agreement instructions given the earlier 40 subjects, was simply to insure that the relative intensity of the descriptive adjectives remained invariant, i.e., was unaffected by the direction of the statement.) An

average rank-order correlation coefficient was then computed to determine the inter-rater reliability.

The instrument used in the experiment proper was a modified Allport-Vernon-Lindzey Scale of Values (1960), containing 60 items. Eighteen different versions, in which the number of alternatives for each item ranged from a 2-point to a 19-point format, were constructed, using the ten adjective descriptors obtained in the first part of the study. The criterion for the construction of each format was that each scale point be approximately equidistant from the ones preceding and following it.

Procedure

The experimenter entered the testing room and proceeded to distribute the rating booklets. Arrangement of the booklets was in such an order that the first subject received a 2-point rating scale, the second received a 3-point rating scale, and so on, until the eighteenth subject received a 19-point rating scale booklet. This procedure was repeated until all subjects had obtained rating booklets. For test-retest purposes, subjects were asked to record their names, course name and number, time and place of meeting, and instructor's name on top of their rating booklets. The subjects were then instructed to open their booklets, read the instructions, record the time, rate the 60 statements, and then record the time at completion of the task. The rating instructions were the same for all the booklets, except that every block of 20 subjects used a different scale to rate the statements. Subjects did not know they were using different rating scales.

After completing the modified Study of Values, the subjects proceeded to fill out an attached criterion measure. Statements in the criterion measure explicitly spelled out what each subscale on the Study of Values was designed to measure, as defined by its test manual. Using a graphic rating scale, each subject was asked to rate the present importance of each of the six value areas in his life.

Three weeks after the first administration, and with the assistance of the identification data provided at the first session, each subject was contacted and received another rating booklet identical to the first. Upon completion, the purpose of the experiment was explained and questions were answered.

Data obtained from the premeasure were analyzed to determine

the internal consistency reliability (Cronbach's alpha, 1951) and concurrent validity. Both measures, pre- and post, were used to assess the test-retest reliability, predictive validity, and the reliability of the criterion measure for attenuation-correction purposes.

A Fisher Z transformation (Fisher, 1921) was undertaken to convert all reliability and validity coefficients in order to insure normality. These transformations were then analyzed by a single classification analysis of variance procedure to determine whether there were significant differences in reliability and validity as a function of rating format. Each of these analyses was segmented by the six value areas in the modified Study of Values.

Following data collection, the responses to each item of the modified Study of Values were converted to dichotomized or trichotomous measures. All even-numbered formats were dichotomized at the center. Responses to the left of center were scored "agree," while those to the right were designated "disagree." The odd-numbered formats were trichotomized, yielding the categories of "agree," "uncertain," and "disagree." The resultant reliability and validity coefficients were then determined for each original and collapsed rating format and subsequently transformed into Fisher Z's. The standard error of the difference between the original and collapsed set of Z's was computed and then divided into the difference between the original and reduced Z coefficients. This procedure, a critical ratio, allowed us to determine whether the original correlations were significantly different from those obtained by collapsing these many-stepped formats to dichotomous or trichotomous measures.

Results

Table 1 summarizes the results of the adjective selection procedure and presents for each statement the proportion of "greater agreement" judgments made by the subjects. Employing the criteria of minimal reversals and approximate equidistance from preceding and succeeding statements, the 10 statements finally selected, together with their scale value, are graphically presented in Figure 1. To ascertain the consistency (inter-rater reliability) with which these statements were ranked, 10 additional subjects proceeded to rank them. An average rank-order correlation coefficient of .99 was obtained, indicating an extremely high degree of

TABLE 1

Scale Values of the Intensity Ratings for the Original Set of Statements

Statement	Proportion of "greater agreement"
I am uncertain	.00
I am uncertain, but probably agree	.08
I hardly agree	.17
I scarcely agree	.20
I minutely agree	.22
I vaguely agree	.29
I barely agree	.30
I slightly agree	.41
I moderately agree	.45
I pretty much agree	.53
I strongly agree	.63
I intensely agree	.74
I immensely agree	.76
I extremely agree	.76
I absolutely agree	.92
I infinitely agree	.94
I unlimitedly agree	.94

agreement among raters as to the rank associated with each statement.

Tables 2 through 5 present the internal-consistency reliability, test-retest reliability, concurrent validity, and predictive validity coefficients (the latter two corrected for criterion attenuation) for each of the 18 rating formats hexacotimized by each of the Allport-Vernon-Lindzey value areas. Table 6 presents the results of analyses

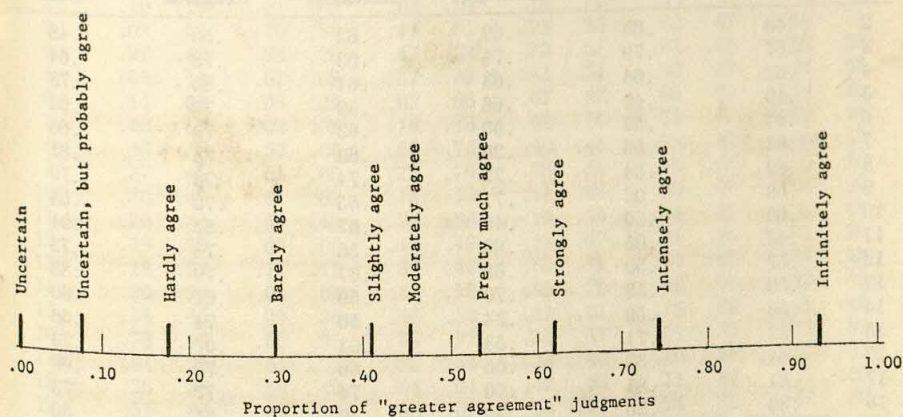


Figure 1. Graphic representation of the scale values of the selected statements.

of variance computed for each value area to assess the extent to which there was a relationship between the rating formats and the reliability and validity measures. This table displays the F ratio for each criterion and value area, indicating whether the relationship found was significant and, if so, to what extent. Examination of Tables 2 through 6, as well as visual inspection of graphs charted from the data contained in these tables, reveals that there is no systematic relationship between predictive validity, concurrent validity, internal-consistency reliability, and test-retest reliability and the number of steps in a Likert-type rating scale. This lack of a systematic relationship was replicated for each of the six value areas encompassed in the modified Allport-Vernon-Lindzey Study of Values.

Table 7 presents the reliability and validity vectors for the 18 original and collapsed rating formats. Figures 2, 3, and 4 graphically display the test-retest reliability, concurrent validity, and predictive validity coefficients for the original and reduced rating formats, respectively. It is apparent that a large degree of overlap exists among each of the three pairs of figures. There appears to be only minimal differences between the reliability and validity vectors based upon the original rating formats and those obtained by collapsing these

TABLE 2

Internal Consistency Reliability Coefficients for Each Rating Format Hexacotomized by Value Area

Format	Theoretical	Political	Economic	Aesthetic	Religious	Social
2	.43	.63	.69	.82	.50	.48
3	.57	.79	.74	.63	.73	.64
4	.62	.64	.63	.61	.85	.73
5	.49	.49	.66	.59	.70	.63
6	.63	.59	.50	.63	.79	.66
7	.63	.56	.26	.63	.88	.81
8	.82	.54	.77	.74	.79	.79
9	.69	.06	.71	.55	.72	.58
10	.66	.50	.46	.67	.83	.91
11	.43	.05	.83	.56	.76	.72
12	.57	.59	.58	.67	.79	.83
13	.50	.53	.70	.59	.61	.60
14	.50	.59	.34	.56	.74	.66
15	.52	.71	.53	.63	.67	.73
16	.64	.52	.66	.66	.70	.69
17	.81	.81	.60	.74	.77	.73
18	.30	.36	.36	.49	.65	.80
19	.62	.24	.69	.64	.79	.87

TABLE 3

Test-Retest Reliability Coefficients for Each Rating Format Hexacotomized by Value Area

Format	Theoretical	Political	Economic	Aesthetic	Religious	Social
2	.64	.99	.99	.99	.99	.98
3	.62	.90	.71	.71	.84	.70
4	.61	.81	.85	.91	.86	.86
5	.78	.81	.63	.87	.89	.83
6	.73	.62	.31	.78	.68	.87
7	.89	.89	.74	.93	.91	.80
8	.92	.81	.83	.94	.88	.88
9	.75	.79	.89	.75	.84	.82
10	.75	.67	.79	.73	.89	.71
11	.15	.76	.86	.89	.82	.84
12	.61	.73	.47	.85	.84	.91
13	.58	.81	.80	.88	.86	.76
14	.47	.65	.58	.71	.78	.79
15	.65	.77	.85	.75	.79	.69
16	.83	.82	.89	.80	.83	.82
17	.64	.75	.85	.61	.69	.82
18	.61	.50	.80	.45	.68	.75
19	.78	.49	.66	.85	.74	.65

TABLE 4

Concurrent Validity Coefficients for Each Rating Format Hexacotomized by Value Area

Format	Theoretical	Concurrent Validity Coefficients										
		Political		Economic		Aesthetic		Religious		Social		
2	.10	.16*	.01	.02*	.03	.04*	.08	.09*	.11	.14*	.43	.62
3	.03	.05	.70	.89	.45	.51	.28	.35	.62	.67	.46	.58
4	.27	.40	.23	.29	.47	.53	.32	.51	.63	.66	.48	.58
5	.05	.13	.07	.08	.37	.39	.45	.54	.86	.87	.52	.60
6	.44	.50	.08	.09	.62	.66	.67	.82	.66	.73	.19	.26
7	.40	.59	.03	.03	.14	.18	.68	.76	.71	.87	.19	.24
8	.43	.52	.57	.60	.65	.75	.40	.43	.78	.81	.50	.58
9	.27	.46	.04	.05	.72	.76	.38	.44	.59	.67	.26	.28
10	.36	.46	.01	.02	.13	.15	.41	.48	.55	.60	.68	.90
11	.26	.36	.39	.43	.72	.86	.19	.35	.64	.76	.33	.41
12	.18	.23	.06	.06	.41	.48	.53	.67	.31	.36	.61	.79
13	.18	.22	.11	.15	.32	.42	.63	.72	.60	.65	.44	.53
14	.20	.24	.04	.06	.14	.16	.55	.75	.62	.66	.15	.18
15	.34	.45	.30	.35	.46	.56	.41	.51	.78	.88	.45	.49
16	.28	.40	.00	.01	.69	.91	.30	.41	.51	.55	.74	.83
17	.81	.93	.54	.72	.33	.51	.33	.40	.16	.17	.22	.27
18	.26	.38	.30	.49	.04	.04	.42	.63	.71	.89	.52	.67
19	.51	.60	.02	.04	.05	.07	.64	.71	.24	.30	.66	.86

* The asterisked columns have been corrected for criterion attenuation.

TABLE 5

Predictive Validity Coefficients for Each Rating Format Hexacotomized by Value Area

Format	Theoretical		Political		Economic		Aesthetic		Religious		Social	
2	.12	.20*	.10	.12*	.01	.01*	.11	.12*	.06	.06*	.50	.73*
3	.10	.13	.54	.68	.55	.62	.49	.62	.49	.54	.07	.08
4	.23	.35	.44	.55	.48	.54	.33	.51	.61	.64	.61	.71
5	.29	.72	.04	.05	.45	.48	.56	.67	.85	.86	.15	.17
6	.39	.44	.10	.11	.55	.59	.61	.74	.75	.83	.11	.15
7	.01	.02	.05	.06	.37	.49	.70	.77	.76	.94	.07	.09
8	.42	.51	.51	.54	.62	.71	.55	.59	.88	.90	.56	.64
9	.05	.09	.04	.06	.81	.84	.44	.51	.58	.66	.20	.22
10	.41	.53	.02	.02	.48	.56	.32	.37	.63	.70	.18	.24
11	.43	.59	.31	.34	.43	.41	.10	.19	.46	.55	.31	.38
12	.52	.66	.07	.08	.40	.46	.42	.54	.43	.50	.24	.31
13	.41	.50	.25	.35	.14	.30	.41	.46	.61	.66	.41	.50
14	.36	.43	.07	.10	.24	.27	.49	.67	.57	.61	.37	.45
15	.22	.29	.36	.41	.64	.77	.34	.43	.61	.69	.43	.47
16	.46	.66	.03	.06	.64	.85	.52	.70	.63	.68	.65	.74
17	.78	.90	.01	.01	.06	.09	.28	.33	.31	.33	.30	.37
18	.24	.34	.18	.29	.03	.04	.36	.54	.44	.55	.39	.50
19	.54	.63	.38	.60	.16	.22	.60	.67	.31	.38	.31	.40

* Corrected for attenuation.

formats to dichotomous and trichotomous measures. Three critical ratios, computed to determine whether these validity and reliability vectors differed, resulted in nonsignificance (Table 8), demonstrating that, regardless of the number of steps originally employed to collect the data, conversion to dichotomous or trichotomous measures does not result in any significant decrement in reliability or validity. Therefore, provided that an adequate number of items are contained on the inventory, increasing the precision of measurement does not eventuate in greater reliability or validity vectors.

Discussion and Conclusions

The evidence from the present study led us to conclude that both reliability and validity are independent of the number of scale points used for Likert-type items. Both internal consistency and stability measures were obtained. The average internal consistency reliability across all areas was .66, while the average test-retest reliability was .82. Both reliability measures, test-retest and internal consistency, were found to be independent of the number of scale points. This finding is consistent with those reported by Bendig (1954), Komorita (1963), Komorita and Graham (1965), and Peabody (1962), contrasts with findings by Symonds (1924) and

TABLE 6
Summary Table of Reliability and Validity Coefficients by Value Area

Criterion	Value Area											
	Theoretical			Political			Economic			Aesthetic		
	F ratio	P	F ratio	F ratio	P	F ratio	F ratio	P	F ratio	F ratio	P	F ratio
Test-Retest	3.74	.005	23.61	.001	18.96	.001	4.82	.001	5.36	.001	7.39	.001
Reliability												
Internal Consistency												
Reliability	2.71	.010	7.66	.001	2.62	.025	0.80	NS	4.32	.001	3.69	.005
Concurrent Validity*	2.71	.010	4.65	.001	4.89	.001	2.87	.01	15.17	.001	12.76	.001
Predictive Validity*	6.11	.001	2.40	.025	12.50	.001	2.58	.025	5.39	.001	1.72	.100

* Corrected for criterion attenuation.

TABLE 7

Reliability and Validity Coefficients for the Original and Reduced Rating Formats

Rating Format	Test-Retest Reliability		Concurrent Validity		Predictive Validity	
	Original Format	Collapsed Format	Original Format	Collapsed Format	Original Format	Collapsed Format
2	.99	.99	.43	.43	.51	.51
3	.70	.70	.47	.47	.07	.07
4	.86	.83	.49	.55	.62	.73
5	.83	.82	.52	.41	.15	.04
6	.88	.80	.19	.28	.12	.19
7	.80	.84	.20	.20	.08	.19
8	.88	.84	.51	.03	.56	.07
9	.82	.78	.26	.42	.21	.22
10	.72	.82	.68	.47	.19	.05
11	.85	.82	.34	.47	.32	.51
12	.92	.88	.62	.64	.24	.27
13	.77	.66	.44	.16	.42	.11
14	.68	.67	.15	.20	.38	.44
15	.70	.65	.45	.40	.44	.37
16	.82	.71	.74	.67	.66	.71
17	.82	.80	.22	.04	.30	.33
18	.75	.62	.52	.36	.39	.40
19	.65	.70	.66	.75	.31	.43

Note.—All values are based upon the social scale.

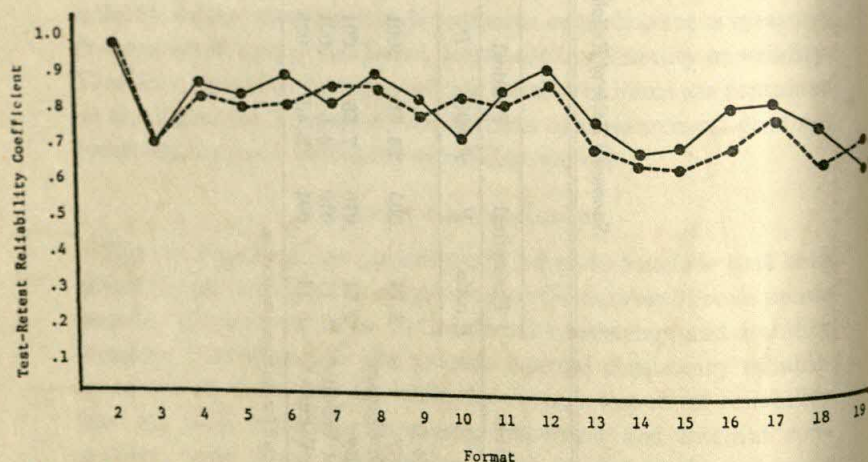


Figure 2. The test-retest reliability coefficients for the original and collapsed rating formats.

* Original Rating Format.

** Collapsed Rating Format.

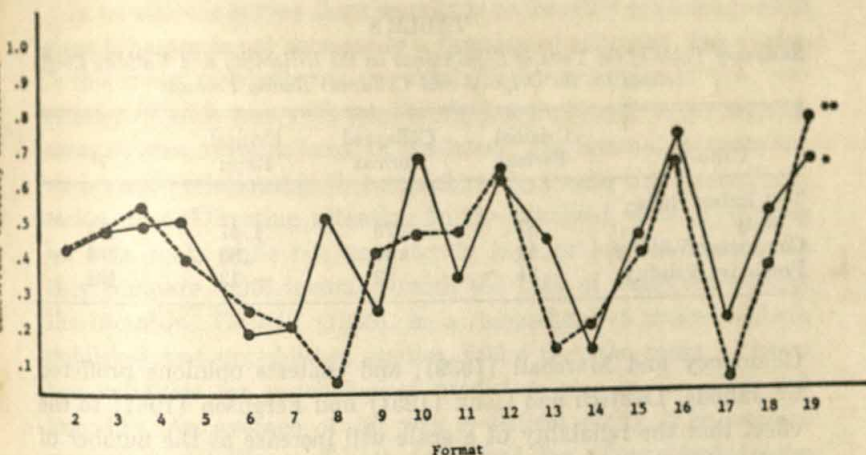


Figure 3. The concurrent validity coefficients for the original and collapsed rating formats.

* Original Rating Format.
 * Collapsed Rating Format.

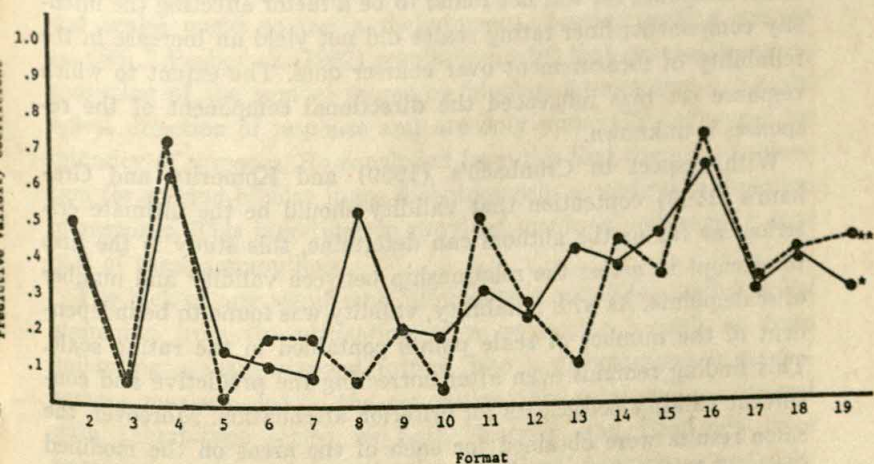


Figure 4. The predictive validity coefficients for the original and collapsed rating formats.

* Original Rating Format.
 * Collapsed Rating Format.

TABLE 8

Summary Table of the Tests of Significance on the Reliability and Validity Coefficients for the Original and Collapsed Rating Formats

Criterion	Original Format	Collapsed Format	Critical Ratio	P
Test Retest Reliability	.82	.78	1.47	NS
Concurrent Validity	.45	.40	.80	NS
Predictive Validity	.34	.33	-.13	NS

Champney and Marshall (1939), and contests opinions proffered by Jahoda, Deutsch and Cook (1951) and Ferguson (1941) to the effect that the reliability of a scale will increase as the number of scale points increase. Based upon the evidence adduced thus far, it would seem that reliability should not be a factor considered in determining Likert-scale rating format, as it is independent of the number of scale steps employed.

Cronbach (1950) claimed that if utilization of finer rating scales, as opposed to coarser ones, increases the reliability of measurement, then this increase should be attributed to the addition of response set variance and not to any increase in the refinement of measurement. Response set was not found to be a factor affecting the intensity component; finer rating scales did not yield an increase in the reliability of measurement over coarser ones. The extent to which response set bias influenced the directional component of the responses is unknown.

With respect to Cronbach's (1959) and Komorita and Gramham's (1965) contention that validity should be the ultimate criterion, as far as the authors can determine, this study is the first to attempt to assess the relationship between validity and number of scale points. As with reliability, validity was found to be independent of the number of scale points contained in the rating scale. This finding remains even after correcting the predictive and concurrent validity coefficients for criterion attenuation. Moreover, the same results were obtained for each of the areas on the modified Study of Values. We can conclude, therefore, that when considering the number of steps to employ in a Likert scale rating format, validity need not be considered because there is no consistent relationship between it and the number of scale steps utilized.

In an attitude survey there usually is no manifest criterion present since behavior is not necessarily a function of attitudes. The choice, in this study, of whether to use either an internal measure (i.e., correlation of each item with total score, less that item) or an external measure was made in favor of the latter. The internal measure has no intrinsic relationship to external reality, while the external criterion does. Directing attention to the obtained validity vectors, we note that, while not consistently high or low, in most cases they compare quite favorably with the bulk of those reported in the literature. Ghiselli (1955), in a comprehensive review of both published and unpublished studies, found that the range of average validities for psychological predictors was in the .30's and low .40's. An average of .50 was a distinct rarity. The average concurrent validity coefficient (corrected for attenuation) in the current study, across all formats and value areas, was .53. The average predictive validity (again corrected for attenuation) was .51.²

Komorita and Graham (1965), in discussing studies by Komorita (1963) and by Bendig (1954), stated that "if it is a valid generalization (i.e., independence of reliability and number of scale steps), the major implication is that, because of simplicity and convenience in administration and scoring, all inventories and scales ought to use a dichotomous, 2-point scoring scheme (p. 989)." Peabody's (1962) results indicated that composite scores, consisting of the sum of scores on bipolar, 6-point scales, mainly reflect direction of response and are only minimally influenced by intensity of response. He concluded from this that there is justification for scoring bipolar items dichotomously according to direction of response. This investigation provided empirical evidence in support of these assumptions.

The lack of any significant differences in reliability and validity stemming from the utilization of a particular format, or from collapsing a many-stepped format into a dichotomous or trichotomous measure, led to the conclusion that total scores obtained with Likert-type scales, as both Peabody and Cronbach have suggested, represent primarily the directional component and only

² Concurrent and predictive validity vectors, uncorrected for criterion attenuation, were .42 and .40, respectively.

to a minor degree the intensity component. Therefore, of the three components contained in a Likert-type composite scale score—direction, intensity, and error—the directional component accounts for the overwhelming majority of the variance.

It has been demonstrated that regardless of the number of steps originally employed to collect the data, conversion of these many-stepped response scales to dichotomous or trichotomous measures does not result in any significant decrement in reliability or validity. Therefore, increasing the precision of measurement does not eventuate in greater reliability or validity vectors, provided that an adequate number of items are contained in the inventory.

One ramification of this finding, if substantiated, would be greater flexibility in the adoption of a given format for a given predictor, criterion, and subject. Since there appears to be independence between reliability and validity vectors and rating format, desirable practical consequences might be obtained from allowing the subject to select the rating format which best suits his needs. This might result in the highly favorable consequence of increasing the subject's motivation to complete the scale. Conversely, if the respondent is not satisfied with a particular rating format, regardless of the reason, the possibility exists that deleterious effects might result from the unsatisfactory rating format-respondent interaction. This interaction could eventuate in a decrement in interest and/or reduced motivation to continue the rating procedure or to complete any remaining parts of the measurement process.

Indeed, it is even conceivable that the subject could record his own responses (open-ended) to each item, without a previously prepared rating format being provided. Subsequently, these subject-produced responses could be transformed to dichotomous or trichotomous measures. Such a strategy might be used to secure the cooperation of individuals who typically are difficult to obtain. By catering to the idiosyncracies of these individuals or, for that matter, any group of respondents, and allowing them to respond in any manner they desire, besides obtaining greater cooperation from these individuals, we might also be able to increase the return rate of our instruments.

A final consideration is the comparison of such data with data

which were previously collected with different rating formats. To overcome this problem, previously collected data could be collapsed into dichotomous or trichotomous measures. This reduction in the precision of measurement, as demonstrated in this research, would not lead to any deleterious effects vis-a-vis reliability or validity. The resultant response distributions, originally based upon different rating formats, could then be directly compared since they would now all be projected from the same base measure.³

A basic question appears to be whether the utilization of fine rating scales increases the refinement of measurement over that which is obtained with coarse dichotomous or trichotomous scales. The overwhelming consistency of results of this study, in addition to those obtained by Peabody (1962), Komorita and Graham (1965), and Bendig (1954), strongly suggests a negative answer to this question.

The primary practical implication of this study is that investigators would be justified in scoring attitude items dichotomously (or trichotomously), according to direction of response, after they have been collected with an instrument that provides for the measurement of the intensity component along with the directional component.

Further research should now be conducted to determine whether the present findings can be generalized beyond the Likert-type scale to different types of scales (e.g., Osgood's Semantic Differential, Thurstone-type scales, graphic rating scales, etc.) and for other purposes (e.g., the rating of behavior, personality, industrial work performance, etc.). It should also be determined whether the conclusions are generalizable to different subject populations defined by such parameters as level of education or ability, and by psychological, experiential, demographic, and ecological characteristics.

REFERENCES

- Allport, G. W., Vernon, P. E., and Lindzey, G. *Study of values*. Boston: Houghton Mifflin Company, 1960.

³ To compare dichotomous and trichotomous measures with each other, the "agree" and "disagree" response categories could be given the weights of one and three, respectively. The remaining "uncertain" response category on the trichotomous format would then be weighted two.

- Bendig, A. W. Reliability and the number of rating scale categories. *Journal of Applied Psychology*, 1954, 38, 38-40.
- Champney, H. and Marshall, H. Optimal refinement of the rating scale. *Journal of Applied Psychology*, 1939, 23, 323-331.
- Cronbach, L. J. Further evidence on response sets and test design. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1950, 10, 3-31.
- Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-334.
- Ferguson, L. W. A study of the Likert technique of attitude scale construction. *Journal of Social Psychology*, 1941, 13, 51-57.
- Fisher, R. A. On the "probable error" of a coefficient of correlation. *Metron*, 1921, 1, Part 4, 1-32.
- Garner, W. R. Rating scales, discriminability, and information transmission. *Psychological Review*, 1960, 67, 343-352.
- Garner, W. R. and Hake, H. W. The amount of information in absolute judgments. *Psychological Review*, 1951, 58, 446-459.
- Ghiselli, E. E. and Brown, C. W. *Personnel and industrial psychology*. New York: McGraw-Hill, 1948.
- Ghiselli, E. E. *The measurement of occupational aptitude*. Berkeley: University of California Press, 1955.
- Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.
- Jahoda, M., Deutsch, M., and Cook, S. W. (Eds.) *Research methods in social relations*. New York: Dryden Press, Inc., 1951.
- Komorita, S. S. Attitude content, intensity, and the neutral point on a Likert scale. *Journal of Social Psychology*, 1963, 61, 327-334.
- Komorita, S. S. and Graham, W. K. Number of scale points and the reliability of scales. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1965, 4, 987-995.
- Likert, R. A technique for the measurement of attitudes. *Archives of Psychology*, 1932, No. 140.
- Peabody, D. Two components in bipolar scales: Direction and extremeness. *Psychological Review*, 1962, 69, 65-73.
- Symonds, P. M. On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology*, 1924, 7, 456-461.

VALIDATION BY THE MULTIGROUP-MULTISCALE MATRIX: AN ADAPTATION OF CAMPBELL AND FISKE'S CONVERGENT AND DISCRIMINANT VALIDATIONAL PROCEDURE

JOHN A. CENTRA

In a well-known paper, Campbell and Fiske (1959) advocate an approach to investigating validity employing a matrix of inter-correlations among tests representing more than one trait, each measured by more than one method. Independent measures of the same trait, they state, should correlate higher with each other than they do with measures of different traits involving the separate methods. In addition these validity values should be higher than the correlations among different traits which happen to employ the same method. Campbell and Fiske refer to the process as convergent (confirmation by independent measurement procedures) and discriminant (distinction of one trait from another) validation by the multitrait-multimethod matrix.

The purpose of this paper is to present an application of the multitrait-multimethod procedure in a different context. While the procedure typically compares independent *methods* of measurement, the application proposed here compares discrete *groups* of individuals; and rather than different individual *traits*, this adaptation compares *scale* scores representing group responses to multidimensional perceptual space. If similar stimulus properties rather than particular subjective interpretations or group biases are being measured, then agreement between the discrete groups in the way they respond to each scale would be expected. As used here, therefore, intergroup agreement on a set of perceptual scales is being assessed, and validity may be defined as the degree of similarity in responses by different constituent groups. The adap-

tation might be referred to as validation by the multigroup-multiscale matrix.

The adaptation will be illustrated in this paper by analyzing group perceptions of their college environment. Multidimensional scales assessing the viewpoints of faculty, administrator, and student groups comprise the multigroup-multiscale matrix.

Method

The environmental assessment instrument used in this study was the Institutional Functioning Inventory (IFI) (Peterson, Centra, Hartnett, and Linn, 1969). Constructed to measure the conditions and emphases at individual colleges and universities, respondents report on what their college is like—what activities go on, how people typically behave—rather than their own behavior or attitudes. The IFI consists of 11 scales, each comprised of 12 items, to which faculty members and administrators on each campus respond; students are judged to be in a position to respond to only the first six scales since they are assumed to be less able to give meaningful responses to the particular items included in the last five scales.

Titles and brief definitions of the IFI scales are given below. More complete scale descriptions as well as information on the development of the IFI may be found in the manual (Peterson et al., 1969).

1. *Intellectual-Aesthetic Extracurriculum* (IAE) refers to the availability of activities for intellectual and aesthetic stimulation outside the classroom.
2. *Freedom* (F) has to do with academic and personal freedom for faculty and students. Low scores suggest an institution that places many restraints on all individuals in the campus community.
3. *Human Diversity* (HD) has to do with the degree to which the faculty and student body are heterogeneous in their backgrounds and present attitudes.
4. *Concern for Improvement of Society* (IS) refers to a desire among people at the institution to apply their knowledge and skills in solving social problems and prompting social change in America.
5. *Concern for Undergraduate Learning* (UL) has to do with

the degree to which the faculty and administration emphasize undergraduate teaching and learning.

6. *Democratic Governance* (DG) has to do with the extent to which individuals in the campus community who are directly affected by a decision have the opportunity to participate in making the decision.
7. *Meeting Local Needs* (MLN) refers to an institutional emphasis on providing educational and cultural opportunities for adults in the surrounding area. High scores indicate availability of adult education, job-related and remedial curricula.
8. *Self-Study and Planning* (SP) has to do with the importance college leaders attach to continuous long-range planning for the total institution.
9. *Concern for Advancing Knowledge* (AK) has to do with the degree to which the institution emphasizes research and scholarship.
10. *Concern for Innovation* (CI) refers to an institutionalized commitment to experimentation with new ideas for educational practice.
11. *Institutional Esprit* (IE) refers to a sense of shared purposes and high morale among faculty and administrators.

Representative samples of faculty and administrators at 22 diversified colleges and universities responded to the IFI during 1968. At 17 of these institutions, a student sample also completed the first six scales of the inventory. The multigroup-multiscale analyses presented in this paper were based on the above sample of institutions. One matrix consists of three groups with six scales ($N = 17$) and the second matrix consists of two groups—faculty and administrators—with all 11 IFI scales ($N = 22$). Group means, compiled with the 12 items in each scale, were intercorrelated. Although the number of institutions was small, it represented several types and sizes of four-year colleges and universities. A larger sample would have been desirable but for the purposes of demonstrating the application presented here, the current sample would appear sufficient.

Internal consistency reliabilities for each scale (coefficient alphas, Cronbach, 1951) were uniformly high (with one exception .86 and higher). These reliabilities, computed for each scale and for

each of the three groups, provided an estimate of reliability based on the mean correlation among items within each scale.

Results

The intercorrelation matrix of faculty, administrator, and student responses to the first six IFI scales is presented in Table 1. The italicized entries in the diagonals, known as the validity values, are particularly important. In the first column, therefore, the validity value of .98 (italicized) is the correlation between faculty and administrator responses to the IAE scale, while in the same column .94 is the correlation between student and administrator responses. The nonitalicized values in the three rectangles represent the correlations of different groups responding to different scales, and with the three triangles are the standard intercorrelations of scales within each group.

According to Campbell and Fiske (1959, pp. 82-83) four aspects of Table 1 bear on the question of validity. First, they recommend that the italicized entries (validity diagonal) be "significantly different from zero and sufficiently large to encourage further examination of validity." With the exception of the DG scale, which has a correlation of .20 between administrator and student responses and a correlation of .30 between faculty and students, this requirement is met (evidence of convergent validity through intergroup agreement).

Second, each validity diagonal value should be higher than the values lying in the column and row of its rectangles; i.e., there should be more agreement among different respondents on the same scale than on different scales. Thus for the IAE scale the correlations of .98 (faculty and administrators), .94 (students and administrators), and .95 (students and faculty) are each higher than correlations for IAE and any other scale (within each of the three rectangles). On the other hand this is not true for the DG scale and in two instances for the F scale. That is, F responses by students correlate .80 with administrator HD responses, which is higher than the .77 correlation between student and administrator F responses; and student F responses correlate .85 with administrator HD responses, which is higher than the .81 correlation between faculty and students on Freedom.

Third, Campbell and Fiske propose that each validity diagonal

TABLE 1

Intercorrelations between Administrator, Faculty and Student Mean Responses to the First Six Scales of the IFI (N = 17 Institutions)

	Administrators						Faculty			Students								
	IAE	F	HD	IS	UL	DG	IAE	F	HD	IS	UL	DG	IAE	F	HD	IS	UL	DG
Administrators	IAE	50																
	F		84															
	HD	47		63														
	IS	52	62		17													
	UL	-05	27	08														
Faculty	DG	33	34	33	72	36												
	IAE	98	44	46	46	-04	29											
	F	47	91	78	46	25	09	45										
	HD	45	80	95	52	-03	15	44										
	IS	61	65	65	92	01	58	56	60									
Students	UL	10	23	06	01	81	09	15	28	-01	04							
	DG	59	40	21	71	37	76	53	30	15	70	23						
	IAE	94	42	44	53	00	27	95	44	41	64	23	53					
	F	33	77	80	53	18	15	31	81	85	66	31	24	39	85			
	HD	31	68	86	43	-06	01	30	71	92	53	04	01	30	30	60		
	IS	49	52	59	84	-01	45	47	48	57	93	06	56	58	64	32	05	18
	UL	01	08	-01	04	71	12	08	16	-06	10	92	24	19	32	05	15	38
	DG	-01	05	02	18	43	20	-02	09	01	30	59	50	15	45	15	38	80

be higher than any of the intercorrelations involving that scale within each group (the values in the triangles). Again, as indicated in Table 1, this requirement is met for each scale except DG and F. For the Freedom scale, for example, faculty responses on F and HD intercorrelate .84, which is higher than the .81 diagonal for F when faculty and student responses are correlated.

The fourth and last aspect recommended by Campbell and Fiske is that the same pattern or ranking of scale (trait) interrelationships be evidenced, regardless of the levels of correlations involved. Inspection of Table 1 suggests that this requirement is fulfilled, with some exceptions on the DG scale.

The first criterion, according to Campbell and Fiske, provides evidence for convergent validity, while the last three criteria are evidence for discriminant validity. On the basis of the multigroup-multiscale matrix, it would appear that the DG scale is somewhat lacking in both convergent and discriminant validity. Specifically, while faculty and administrators agree fairly substantially about the extent of Democratic Governance on their campus (.76), students disagree with both adult groups. In addition, the DG scale correlates highly with other scales within each group: .72 with IS for administrators, .70 with IS for faculty, and .80 with UL for students.

The F and HD scales, according to the multigroup-multiscale analysis, intercorrelate highly within each of the response groups and thus are somewhat lacking in discriminant validity. In spite of this, it would appear that these two scales are still valid for faculty and administrators (both with values around .90 in the validity), and only less so for students (evidence of convergent validity). In fact, for students, only the F scale's validity diagonal value (.81) failed to meet the Campbell and Fiske criteria.

Faculty-Administrator Response Comparisons

Additional validity analyses by the multigroup-multiscale method are presented in Table 2, in which faculty and administrator responses at 22 institutions are compared. Among the last five scales, the validity diagonal values of .96, .81, .96, and .64 for MLN, SP, AK, and IE respectively are each higher than values in their respective rows and columns. These scales, therefore, appear to meet several of the validity criteria. For the Concern

TABLE 2
Intercorrelations between Administrator and Faculty Mean Responses to the IFI Scales
(*N* = 28 Institutions)

Administrators														Faculty													
IAE	F	HD	IS	UL	DG	MLN	SP	AK	CI	IE	IAE	F	HD	IS	UL	DG	MLN	SP	AK	CI	IE						
IAE	51																										
F	45	83																									
HD	47	67	63																								
IS	-30	03	-03	-01																							
UL	32	32	31	65	24																						
DG	-20	04	-21	04	-21	07																					
MLN	-05	-23	-10	13	23	42	34																				
SP	61	47	48	59	-61	24	00	-26																			
AK	00	00	49	49	62	12	68	-12																			
CI	25	-47	-28	12	24	22	42	65	-27	36																	
IE																											
IAE	98	46	45	43	-25	28	-26	-09	55	-04	-26																
F	47	92	79	52	04	09	-23	-38	35	13	-53	47															
HD	42	79	94	55	-18	12	-17	-19	50	23	-36	44	84														
IS	47	68	63	92	-13	48	-10	-10	58	25	-29	46	60	63													
UL	-24	03	-03	-07	89	07	-41	-02	-59	24	-01	-16	08	-15	-05												
DG	47	39	21	65	29	74	-09	12	26	31	05	44	31	13	60	24											
MLN	-24	-16	-20	08	-17	12	96	33	-04	15	48	-31	-20	-16	-05	-37	-04										
SP	02	-32	-27	06	36	40	14	81	-42	52	48	05	-36	-31	07	21	32	15									
AK	61	53	54	58	-68	17	03	31	96	-17	-37	57	45	60	64	-62	22	60	-43								
CI	28	60	66	79	28	60	-17	11	21	69	-20	32	53	54	73	22	61	-12	29	24							
IE	-05	-47	-39	00	12	03	20	34	-12	-05	64	-02	-44	-38	-01	15	25	77	41	-18	-16						

for Innovation scale (CI), however, faculty and administrator responses correlate .59. While this correlation is significantly greater than zero, four scales in the CI row have higher correlations. The CI scale, therefore, lacks both convergent and discriminant validity according to the Campbell and Fiske criteria. The remaining 10 scales, however, generally fulfill the four criteria.

Discussion

Interpretation of the multigroup-multiscale matrix differs considerably from that of the multitrait-multimethod matrix. According to the latter, assessment methods converge (methodological triangulation) to confirm the measurement of a trait; according to the multigroup analysis, groups converge or agree in their assessment (perceptions) or stimulus properties (scales). If, as in this study, discrete groups respond similarly to scales meant to assess the environment they inhabit, then it becomes more reasonable to assume that the scales are measuring characteristics or conditions of that environment. If, on the other hand, there is lack of agreement between groups, environmental features (scales) are less objectively measured than they might be. In short, the scales more likely reflect subjective interpretations of the environment rather than relatively objective environmental constructs. Of course the logic of the multigroup approach assumes that each group is fully acquainted with the perceptual space (environment, in this study) represented by the scales. Otherwise lack of group agreement could simply indicate lack of knowledge by one group.

In this study faculty and administrative groups agreed in their responses to 10 of the 11 environmental scales. For these 10 scales, therefore, similar institutional functions are being assessed regardless of whether the respondent group consists of faculty or administrators. The Concern for Innovation (CI) scale, on the other hand, which lacked convergent (agreement) validity, would appear to measure somewhat different institutional emphases, depending on whether faculty or administrators were responding. Similarly the DG and F scales apparently assess different institutional functions when students respond than when faculty or administrators respond. The existence of Democratic Governance

and Freedom on campus, in other words, means something quite different to students than to either adult group.

It is important to note, however, that the multigroup-multiscale adaptation does not confirm the theory underlying the measurement. As with the multitrait-multimethod procedure, the existence of a construct is established less by convergent validity than by predicting correlations with other measures or by other strategies of construct validity (APA, 1966; Cronbach and Meehl, 1955). Convergent and discriminant validity information through the multigroup-multiscale matrix does, nevertheless, appear to provide useful insights into how a current instrument is functioning and might be improved. For example, in this study the Concern for Innovation scale (CI) probably needs to be more clearly delineated in view of its overlap with other scales and its relatively low intergroup agreement. Changes made in the scale would hopefully lead to improving its subsequent "validity coefficient"; in this manner, as Campbell and Fiske suggest, validation may be viewed as an ongoing program for improving measurement devices.

REFERENCES

- American Psychological Association Standards for Educational and Psychological Tests and Manuals. APA, Washington, D. C., 1966.
- Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-334.
- Cronbach, L. J. and Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, 52, 281-302.
- Peterson, R. E., Centra, J. A., Hartnett, R. T., and Linn, R. L. *Institutional Functioning Inventory*. Technical Manual. Princeton, N. J.: Educational Testing Service, 1969.

THE ROBUSTNESS OF TILTON'S MEASURE OF OVERLAP¹

RICHARD S. ELSTER

U. S. Naval Post Graduate School
Monterey, California

MARVIN D. DUNNETTE

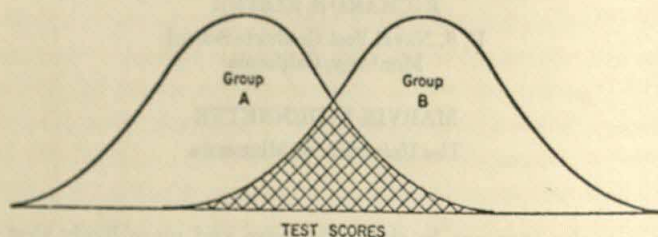
The University of Minnesota

As samples increase in size, it is more and more likely that any given null hypothesis will be rejected at a statistically significant level. Many authors have discussed the implications of this fact (Baker, 1966; Lykken, 1968; Nunnally, 1960; Rozeboom, 1960) and have severely criticized the continuing use by psychologists of the statistical reject-accept hypothesis testing model. The essence of their criticisms is that when the model is used by psychologists, (e.g. by applying F or t statistics to examine mean differences), large samples are likely to yield "significant" statistics even when the magnitude of mean differences is trivial. Thus, scientific conclusions derived from such strategies may be erected on an accumulating array of triviality (Dunnette, 1966). The problem has been stated forcefully by Hays, 1965; p. 326: "Virtually any study can be made to show significant results if one uses enough subjects, regardless of how nonsensical the content may be."

It is agreed by most such critics that psychologists should make greater use of measures expressing the practical importance of dif-

¹ The research described here was completed by the first author as his dissertation research for the PhD degree at the University of Minnesota. The authors appreciate the support given to Elster during his graduate studies by a pre-doctoral fellowship from the National Institute of Health. Expenses related to this research were also defrayed in part by a Behavioral Science Research Grant to Dunnette from the General Electric Foundation.

ferences instead of relying entirely on the hypothesis testing models. One such statistic (Tilton, 1937) expresses the degree of overlap between two distributions. The overlap (O) between score distributions obtained by two groups is defined as the percentage of persons in one of the groups whose scores may be matched by persons in the second group. In the diagram below, the shaded area designates the overlapping portion of two distributions. Two groups which overlap completely will have identical scores and an O of 100 per cent. Two groups which are entirely separated will, of



course, have different means; but, more important, the O will be 0 per cent. Values of O between 0% and 100% give meaningful and easily interpretable estimates of the practical importance of a mean difference. Tilton's statistic is based on the ratio of the difference between the means of the two groups to the average of the two standard deviations. Table 1 (from Tilton, 1937, p. 658) shows the relevant information for estimating the value of O . The table was developed by assuming that the two sample distributions of scores are random samples from normally distributed populations with the same standard deviations, assumptions which Guion (1965) claims are rather rarely fulfilled when working in an applied situation.

We agree with Guion that it is rare indeed for scores in two *samples* to be perfectly distributed normally or to have identical standard deviations. However, we must disagree with him about the presumed rarity of normality in population distributions; and, of course, the equivalence of the population standard deviations can be estimated by computing confidence intervals based on the sample statistics.

Even so, it is important to examine the effects on the value of O of non-normality in the population distributions and of departures

TABLE 1

Tilton's Index and Tilton's Percentage Overlapping D/SD
 $\text{ave.} = |\bar{X}_1 - \bar{X}_2| / (SD_1 + SD_2) / 2$

<i>D/SD ave.</i>	<i>% Overlap (O)</i>	<i>D/SD ave.</i>	<i>% Overlap (O)</i>
0.000	100	1.349	50
0.025	99	1.381	49
0.050	98	1.413	48
0.075	97	1.445	47
0.100	96	1.478	46
0.125	95	1.511	45
0.151	94	1.544	44
0.176	93	1.578	43
0.201	92	1.613	42
0.226	91	1.648	41
0.251	90	1.683	40
0.277	89	1.719	39
0.302	88	1.756	38
0.327	87	1.793	37
0.353	86	1.831	36
0.378	85	1.869	35
0.403	84	1.908	34
0.429	83	1.948	33
0.455	82	1.989	32
0.481	81	2.030	31
0.507	80	2.073	30
0.533	79	2.116	29
0.559	78	2.161	28
0.585	77	2.206	27
0.611	76	2.253	26
0.637	75	2.301	25
0.664	74	2.350	24
0.690	73	2.401	23
0.717	72	2.453	22
0.744	71	2.507	21
0.771	70	2.563	20
0.798	69	2.621	19
0.825	68	2.682	18
0.852	67	2.744	17
0.880	66	2.810	16
0.908	65	2.879	15
0.935	64	2.952	14
0.963	63	3.028	13
0.992	62	3.110	12
1.020	61	3.196	11
1.049	60	3.290	10
1.078	59	3.391	9
1.107	58	3.501	8
1.136	57	3.624	7
1.166	56	3.762	6
1.196	55	3.920	5
1.226	54	4.107	4
1.256	53	4.340	3
1.287	52	4.653	2
1.318	51	5.152	1

from equality in population variances. In other words, how robust is Tilton's overlap statistic?

Method

In order to investigate the robustness of O , we used a digital computer to generate pairs of populations of predetermined size and shape with specified means and variances. Fifty samples of specified size were then selected randomly from each pair of populations. (After a pair of samples had been drawn and Tilton's O calculated, their "scores" were returned to the populations.) Descriptive statistics were computed for the distribution of 50 sample Tilton overlap values. The mean and median sample values were then compared with the actual overlap between the two populations. The actual population overlap was obtained simply by counting the number of scores in one population that could be matched by scores in the other population and converting this count to a percentage of each population's N .

All the populations developed consisted of 2000 whole numbers. Normal curves were generated by using a computer program which produced psuedo-random numbers. This generator "drew" numbers from a normal distribution having a mean of zero and a variance of one. The random numbers were then transformed to the desired mean and variance by a simple linear transformation. Non-normal distributions were generated by using a computer program that allowed the user a great deal of freedom in establishing the shapes and central tendencies of the distributions to be used. The only two limitations in forming non-normal distributions were that the distribution could have no more than 100 unique score values and the total number of observations could be no larger than 2000.

In addition to calculating the mean and variance of each population distribution generated, measures of skewness and kurtosis were computed. The method used to calculate the measures of kurtosis and skewness were those given in McNemar (1962, pp. 25-28, and pp. 78-79). When the population's skewness index is positive, its distribution is skewed to the right; a negative index indicates a skewed-left distribution. When the population's kurtosis index is less than zero, its distribution is somewhat flat-topped; when the kurtosis index is greater than zero, the distribution is peaked with a higher tail than those of a normal distribu-

tion. (Examples of distributions with their kurtosis or skewness indexes can be found in McNemar, 1962, p. 27, and in Ghiselli, 1964, p. 58.)

Analyses and Results

The analyses consisted of two phases. First, the effect of unequal population variances upon Tilton's overlap values was examined. Second, the impact of non-normal population distributions upon sample Tilton's overlap values was studied.

Unequal Population Variances

The first analyses were intended to investigate the effects of unequal population variances upon estimates of population overlap calculated using Tilton's method with sample data. For each analysis, a pair of normal populations was generated with a desired mean difference and a specified relationship between their variances. One population of each pair always had a variance near 100 while the other population's variance was made some desired multiple of 100. If, for instance, the first distribution had a variance of 100, and the second a variance of 121, the populations were said to have a variance ratio of 1.21:1. The other variance ratios used were 1:1, 1.44:1, 2.25:1, 3.9:1, and 8.9:1.

Normal populations having the six population variance ratios were then generated having each of these five population mean differences: 0, 5, 10, 20, and 30. Fifty pairs of random samples of 100 observations each were then drawn from each of the 30 population variance-ratio \times mean-difference combinations, and the mean sample Tilton's O calculated for each set of 50 values. The value of the population overlap, determined by calculating the number of scores in one population that could be matched by scores in the other population and dividing this total by the size of each population ($N = 2000$), was then subtracted from the mean sample Tilton's overlap. These differences were then used to assess how well the Tilton's overlap values approximated the overlaps of the populations from which the samples were drawn. Figure 1 shows the findings of the investigations concerning the effects of population heteroscedasticity and varying population mean differences.

Two major conclusions can be drawn from the information shown in Figure 1: (a) at any particular population mean differ-

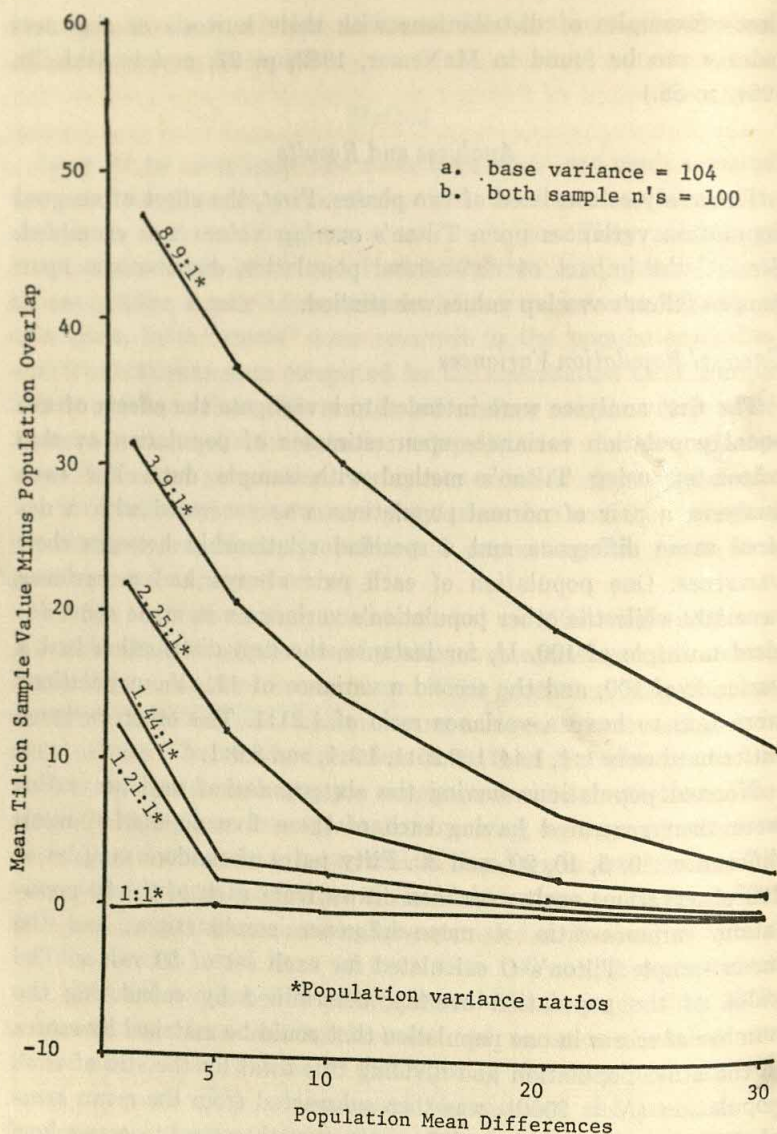


Figure 1. Value of the mean sample Tilton's overlap minus the population overlap for pairs of normal populations a, b.

ence, the greater the violation of the assumption of equal variances, the greater is the deviation of the typical sample Tilton's overlap value from the population overlap figure; (b) the effects of any particular violation of the assumption of equal population vari-

ance decreases as the population mean difference increases. It is also somewhat reassuring to note that the errors shown by the deviations graphed in Figure 1 are all in the conservative directions—the Tilton's; that is, overlap values obtained from samples tend to overestimate the overlap of the populations from which they are drawn.

Effects of Non-normal Population Distribution

The effect of distribution shape upon Tilton's overlap as an estimator of population overlap was investigated by using non-normal distributions established by the investigators. Table 2 contains the descriptive statistics of the eight non-normal distributions which were used. Three of the distributions are skewed right, three skewed left, one rectangular, and one bi-modal. Each of these distributions contained 2000 "scores."

With eight non-normal distributions and the various normal distributions possible, a large number of potential pairings were available. This number was increased further by the variety of population mean differences which were possible. The decision was made to pair first each of the non-normal distributions with a normal distribution of the same size (N) and the same variance. Using distributions of equal N s and variances controlled these variables so that the effects of varying shapes could be seen alone.

In every pairing of a non-normal with a normal population, the normally distributed population was established having a mean higher than that of the non-normal population. This fact is particularly important to remember when examining the analyses in-

TABLE 2

The Symmetry Statistic and the Mean and Variances of the Eight Non-normal Distributions

Distribution Type	Mean	Variance	Symmetry Statistic
Skewed Right (SR_1)	44.47	51.18	.48
Skewed Right (SR_2)	41.61	51.34	.77
Skewed Right (SR_3)	42.13	51.00	.70
Skewed Left (SL_1)	48.53	51.18	-.48
Skewed Left (SL_2)	48.43	51.59	-.76
Skewed Left (SL_3)	49.97	51.00	-.70
Rectangular	62.00	52.05	0.00
Bi-Modal	47.99	53.92	.07

cluding a skewed population. A skewed-right population would intersect the normal population with its skewed tail. A skewed-left population would have the larger part of its distribution overlapping with the normal population.

The sampling distributions of Tilton's overlap values were formed by taking 50 pairs of samples from the populations involved in each analysis. All samples had n s of 100. The mean Tilton's overlap values were used in describing the central tendencies of the sampling distributions. Table 3 includes the population overlaps and the mean sample Tilton's overlap for each of the population pairings investigated.

The data in Table 3 show rather remarkable agreements between the population overlaps and the mean Tilton's overlap sample values. The exceptions to this statement are all found in the cases where populations were established with a mean difference

TABLE 3

Values of the Population Overlap and the Mean Tilton Overlaps from Analyses Involving Non-normal Populations

Distributions Paired	Population Mean Difference	Population Overlap	Mean Sample Tilton's
Skewed Right	0	85.00	93.38
(SR ₁) and Normal	5	67.80	70.90
	10	47.40	47.04
Skewed Right	0	78.00	99.60
(SR ₂) and Normal	5	65.10	72.96
	10	46.00	49.02
Skewed Right	0	81.50	99.66
(SR ₃) and Normal	5	66.90	72.50
	10	46.40	48.70
Skewed Left	0	86.40	99.28
(SL ₁) and Normal	5	75.80	71.54
	10	51.20	47.54
Skewed Left	0	80.00	99.72
(SL ₂) and Normal	5	73.40	71.64
	10	49.30	47.76
Skewed Left	0	82.70	99.86
(SL ₃) and Normal	5	75.80	72.20
	10	50.20	48.24
Rectangular and Normal	0	80.50	99.00
	5	74.30	71.76
	10	54.09	47.62
Bi-Model and Normal	0	68.70	98.66
	5	63.50	71.58
	10	51.00	48.28

* Both sample n s = 100.

of zero. The sample Tilton's values in these cases are too high because of the variation of the sample mean differences around the population mean difference of zero. When Tilton's index is near zero, the corresponding value of Tilton's overlap is near 100 per cent. It is somewhat paradoxical that the Tilton's overlap estimate obtained when the populations have a mean difference of zero seems to make "more practical sense" than the count overlap on the same populations. Results from Tilton's technique make more practical sense because they yield a value near 100 per cent while the count overlap takes on values of less than 100 per cent—the amount that it is less depending upon the relationships between the population's N s, variances, and "shapes."

The data in Table 3 also indicate that for the skewed-right distributions the differences between the typical Tilton's overlap sample values and the population overlaps increase as the populations become increasingly skewed. For the skewed-left populations, the same differences decrease as the skewness increases. The overlaps of the rectangular and normal populations were always underestimated by the sample overlaps. The direction of the error when bi-modal populations were used depended upon the difference between the means of the two populations.

Effects of Two Non-normal Population Distributions

From the many combinations of non-normal population distributions that could have been formed, only six pairings were investigated. The means, variances, and symmetry statistics of the population distributions which were used are given in Table 2.

The accuracy of the mean sample Tilton's overlaps in Table 4 is somewhat surprising. The most inaccurate results were obtained from the pairing of two populations which had bi-modal distributions. The mean Tilton's sample overlap is 15 per cent greater than the population overlap; here again the error is in the conservative direction. When both the magnitude and the direction of the errors are considered, the pairing of two rectangular populations causes the most important error in the overlap estimates given by Tilton's method. In this case, the mean Tilton's sample overlap is 11.9 per cent below the overlap of the two populations' distributions.

These writers by no means claim that the analyses shown in

TABLE 4

Values of the Population Overlap, the Mean Tilton Overlap, and the Standard Error of Tilton's Overlap from Analyses Involving Pairs of Non-Normal Populations ^{a, b}

Distributions Paired	Population Mean Difference	Population Overlap (%)	Mean Sample Tilton's (%)	Standard Error of Tilton's Overlap
SR_1 and Rectangular	19.87	17.2	16.12	2.946
SL_1 and Rectangular	12.13	42.9	68.64	3.840
SR_2 and SL_2	7.74	54.5	56.60	6.200
Bi-Modal and Bi-Modal	10.00	34.0	49.00	4.087
Bi-Modal and Rectangular	14.00	37.5	32.38	3.678
Rectangular and Rectangular	10.00	60.0	48.10	4.360

^a Both sample n s = 100.

^b Both population N s = 2000.

Table 4 fully illuminate the effects of non-normal population distributions. It is hoped, however, that the variety of pairings used in these analyses will at least let the reader form a general impression of the impact of non-normality upon Tilton's overlap values obtained for samples from non-normal populations having nearly equal variances.

Effects of Non-normal, Heteroscedastic Population Distributions

Six analyses were run in an effort to throw a small ray of light on the effects of having pairs of non-normal populations with unequal variances. Table 5 includes three pairings of non-normal populations having equal variances and the same three pairings when the populations had unequal variances.

The results in Table 5 are quite uniform: when the population variances are extremely different, the typical Tilton's sample overlap values greatly overestimate the population overlaps. When the population variances are made widely different, the differences between the mean Tilton's sample values and the population overlaps all increase and change in direction from underestimation to overestimation. The effects of unequal variances when the populations are non-normal is very similar to the results found with normal populations having unequal variances shown in Figure 1.

TABLE 5

*Analyses of the Effects of Non-normality and Heteroscedasticity—All Population $N_s = 2000$
and Both Sample $N_s = 100$*

Distributions Paired	Population Mean Difference	Population Variance	Population Overlap (%)	Mean Sample Tilton's	Standard Error of Tilton's
SL_4 and Rectangular	12.13	51.00	42.9	38.64	3.840
SL_4 and Rectangular	12.63	51.00	46.2	55.08	4.087
Bi-Modal and Rectangular	14.00	208.46	37.5	32.28	3.678
Bi-Modal and Rectangular	14.51	53.92	41.1	50.18	4.617
Rectangular and Rectangular	10.00	53.05	60.0	48.10	4.360
Rectangular and Rectangular	9.50	52.05	50.0	65.30	4.834
Rectangular and Rectangular		52.05			
Rectangular and Rectangular		208.46			

As can be seen in both Figure 1 and Table 5, the errors (differences between the mean sample Tilton's overlap and the populations' overlap) are in the conservative direction. The practitioner can feel reasonably sure that using Tilton's method with samples from populations having unequal variances, whether the populations are normal or non-normal, will not result in an underestimate of the overlap existing in the parent populations.

Summary

Pairs of populations were established having designated N_s , variances, means, and distribution "shapes." Fifty pairs of random samples of some desired size were then taken from the pair of populations. Descriptive statistics were computed for the distribution of 50 sample Tilton overlap values. The mean sample values were then compared with the count overlap of the two populations. The count overlap was obtained by calculating the number of scores in one population that could be matched in the other population and converting this count to a percentage of the two populations' N_s .

The results show the typical sample Tilton overlaps accurately reflect the population overlap under most violations of the assumptions underlying Tilton's statistic. However, unequal population variances, when the populations are both normal, cause the Tilton sample values to greatly overestimate the population overlap. This error does decrease as the populations' mean differences increase. When the populations having unequal variances are both non-normal, the Tilton sample values still overestimate the population overlap. Tilton's overlap, as was mentioned earlier, may be considered to be more meaningful than a count overlap when the populations have unequal variances. The separation of populations having unequal variances seems to be better represented by Tilton's technique than it is by a count overlap measure.

The effects of distribution shape are expectedly variable. When only one of the populations in the pair is normally distributed, and has a standard deviation nearly equal to that of the non-normal distribution, the typical sample Tilton estimates of the population overlap are all quite accurate. If both populations in the pair are non-normal in form, the magnitude and the direction of the error depends upon the shapes of the curves used in the analysis. Using a pair of bi-modal populations results in the greatest discrepancy between the typical Tilton sample value and the population overlap. The error in this case is again in the conservative direction—the sample Tilton overlaps overestimating the population overlap.

The practitioner will probably wish to pay greatest heed to the effects of unequal population variances. With normal distributions, sample Tilton's overlaps will substantially overestimate the population overlap unless the two populations have unequal means. As the populations' variances become more unequal this overestimation becomes greater at any particular population mean difference. The three analyses using heteroscedastic non-normal populations also show the sample Tilton overlaps to overestimate the population overlap. The direction of this error may cause the user to reject some measures, but it should not lead him to choose any bad ones. If an investigator's purpose is to select a measure that effectively separates two groups, then the Tilton estimate probably reflects quite accurately the conservatism that should be a part of such decisions.

REFERENCES

- Baker, David. The test of significance in psychological research. *Psychological Bulletin*, 1966, 66, 423-438.
- Dunnette, M. D. Fads, fashions, and folderol in psychology. *American Psychologist*, 1966, 21, 343-352.
- Ghiselli, E. E. *Theory of psychological measurement*. New York: McGraw-Hill, 1964.
- Guion, R. M. *Personnel testing*. New York: McGraw-Hill, 1965.
- Hays, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.
- Lykken, D. T. Statistical significance in psychological research. *Psychological Bulletin*, 1968, 70, 151-160.
- McNemar, Quinn. *Psychological statistics*. New York: Wiley, 1962.
- Nunnally, Jum, The place of statistics in psychology. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1960, 20, 641-648.
- Rozeboom, W. W. The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 1960, 67, 416-428.
- Tilton, J. W. The measurement of overlapping. *Journal of Educational Psychology*, 1937, 28, 656-662.

THE RELATIVE EFFICIENCY OF REGRESSION AND SIMPLE UNIT PREDICTOR WEIGHTS IN APPLIED DIFFERENTIAL PSYCHOLOGY¹

FRANK L. SCHMIDT
Michigan State University

A very common problem in the behavioral and social sciences is the prediction of the standing of a person or thing on one variable, usually designated the criterion, from his or its standing on a number of other variables, usually called the predictors. Least-squared error multiple regression weights are most commonly used in weighting the predictors into a composite. These weights, which minimize, over the cases in the sample, the sum of the squared deviations of the observed from the predicted criterion score, are calculated from the normal equations which express the minimization conditions (Anderson, 1958). The fitting of the regression weights to the idiosyncracies of the initial sample leads to a decrease in effectiveness when these weights are applied to a new sample in which these particular idiosyncracies are not present. This "shrinkage" is often substantial in practical situations (e.g., see: Kurtz, 1948; Cureton, 1950; and Kirkpatrick, 1951), especially when the initial sample is small. And small samples, as Lawshe and Schucker (1959) point out, are the rule rather than the exception in many areas of applied psychology.

Certain other approaches to the weighting problem produce

¹ This study is based on a dissertation submitted in partial fulfillment of the requirements for the PhD degree at Purdue University. The author is indebted to Professors Hubert E. Brogden, Joseph Tiffin, and Norville M. Downie for their assistance and advice and to Dr. Vernon Urry for writing the necessary computer program.

weights independent of information in the first sample. Raw predictor scores may be summed to yield a composite in which each test is weighted by its standard deviation. Or, weights may be derived a priori from a theory. Another approach is to standardize each predictor and then sum predictor scores into a composite, thus weighting all predictors equally. In the present study, the latter approach to weighting was compared to multiple regression methodology in the data domain of applied differential psychology. Because of the many socially and economically important decisions made about individuals and groups of individuals on the basis of psychological measurements of various kinds, the weighting problem is perhaps of more practical significance in this data domain than in many other areas of the behavioral sciences.

Of previous studies, the most significant is probably that of Lawshe and Schucker (1959). These researchers examined the relative efficiency of four test weighting methods: (1) the simple addition of raw scores (which weighted each test by its SD); (2) weighting of raw scores by the test SD (which weighted each test by its variance); (3) weighting of raw scores by $1/SD$ of the test; and (4) least squares multiple regression weights. Three batteries of three tests each were used, each with a different average predictor intercorrelation (.23, .37, and .61); average validity was held constant across batteries. The criterion was a dichotomized GPA measure. Regression weights were derived on samples of 20, 40, and 90, and all equations were applied to two hold-out samples of 75 each. None of the weighting methods showed any advantage over any of the other methods. Also, regression weights derived on the larger samples did not show superiority to those derived on the sample of 20.

Trattner's (1963) findings with respect to different predictor weighting methods are confounded with the effects of different test selection methods used. It should be noted that the problem as defined in the present study assumes no a posteriori selection of predictors. The assumption is that either all predictors selected on an a priori basis are used in the final equation or that selection among the various predictors is independent of their performance in the first sample. Using a first sample of approximately 100 and cross-validation samples ranging from 60 to 125, Trattner found that none of the four combinations of test weighting and selection

methods that he examined showed any advantage over the other combinations.

Wesman and Bennett (1959), using N s of 262 to 449, found no advantage for least squares weights over a simple summing of raw scores in the prediction of grade point average from the three subtests of the College Qualification Test. Grant and Bray (1970), with N s in the neighborhood of 200, found that the decrease in correlation in the cross-validation sample resulting from the use of unit rather than regression weights was only .01.

Researchers concerned with the effects of item weighting within tests have almost invariably concluded that little or nothing is to be gained by such weighting under most conditions encountered in practice (e.g., see: Guilford, Lovell and Williams, 1942; Harper and Dunlap, 1942; Phillips, 1943).

Two other studies not directly concerned with predictor weighting have produced additional evidence for questioning the utility of differential weights (Ryans, 1954; Ewen, 1956).

There is at least one rather basic criticism that can be leveled at all of these studies: they have all compared the performance of different weighting methods *in a new sample*. What is really needed is information on the relative performance of the different methods when sampling is not a factor, i.e., information on their performance in the population. When the applied psychologist is faced, for example, with the task of deciding between two different sets of weights to be applied to a given battery of tests to predict a criterion of, say, job success, the information he needs to make this decision is not how the two sets of weights compare in some new sample of 30, 80, or 100, but rather how their performances compare in the long run, i.e., in the population of potential applicants for the job. While it is true that a random sample can provide an unbiased estimate of the relative performances of different sets of weights in the population, sampling error may erase or even reverse in the sample the actual inferiority-superiority relationship existing in the population. In the Lawshe and Schucker (1959) study, for example, cross-validation on samples of 75 revealed no difference in performance between composites produced by weighting each test by its SD and those produced when each test was weighted by $1/SD$. It is safe to speculate that these two sets of weights, correlated -1.00 , differ in their efficiency in the popu-

lation. Brogden (1946), among others, has shown that, in certain situations, even small increments in the correlation coefficient can be of significant practical import.

Definitions of symbols used in this study are as follows:

- p = the number of predictors being used.
- N = sample size.
- Σ_{xx} = the $(p + 1) \times (p + 1)$ population correlation matrix.
- Σ_x = the $p \times p$ population correlation matrix of the p predictors.
- Σ_x = the $p \times 1$ population vector of predictor validities.
- R = a $(p + 1) \times (p + 1)$ sample correlation matrix.
- R^* = a $p \times p$ sample correlation matrix of the p predictors.
- $\rho(\underline{\beta})$ = the population multiple correlation produced by the actual, infallible population weights, $\underline{\beta}$.
- $\rho(\underline{\hat{\beta}}_i)$ = a population correlation produced by a set of fallible regression weights, $\underline{\hat{\beta}}_i$, computed on a sample from the population.
- $\rho(\underline{1})$ = the population correlation produced when unit weights are applied to the predictors.

Finally, the symbol ϵ is used to indicate the operation of taking the expected value.

Figure 1 shows that, for any fixed N , p and Σ_{xx} , $\rho(\underline{\hat{\beta}})$ forms a distribution in the population with some variance, $\sigma^2(\underline{\hat{\beta}})$. $\rho(\underline{\beta})$ and $\rho(\underline{1})$, on the other hand, are point distributions, each with zero variance. $\underline{\beta}$ and $\underline{1}$ are constant in value and, whenever applied to the population, yield $\rho(\underline{1})$ and $\rho(\underline{\beta})$, respectively, with a probability of 1.00. Each $\rho(\underline{\hat{\beta}}_i)$, in contrast is produced by a somewhat different set of weights, depending on the sampling error of the particular sample that happens to be drawn. Thus if one draws a large number of samples, each of size N , and computes a $\underline{\hat{\beta}}_i$ for each sample, the resulting $\rho(\underline{\hat{\beta}}_i)$ coefficients will form a distribution more or less similar to that shown in Figure 1. It should be noted that $\rho(\underline{\hat{\beta}}_i)$ can be negative if the weights computed on sample i are particularly inefficient. Point c on the abscissa represents $\epsilon\rho(\underline{\hat{\beta}})$, and the distance ac represents the average superiority in the population of regression over unit weights for a given N , p , and Σ_{xx} . It is this distance that the present study was designed to estimate.

Consideration of possible approaches to this problem led to the selection of a Monte Carlo methodology. Because the equation for

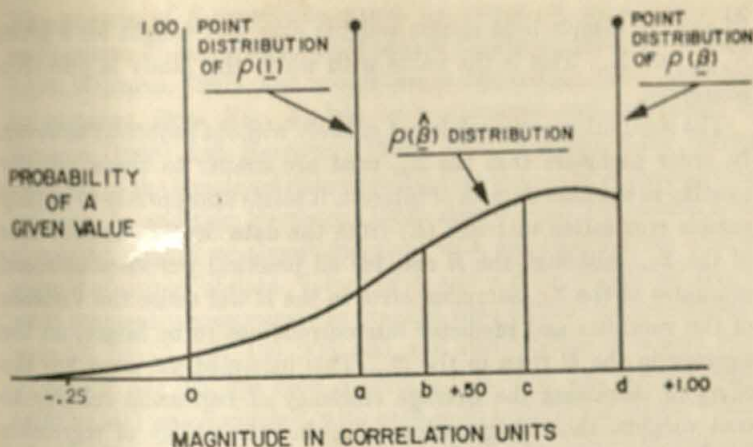


Figure 1. Depiction of the distributions of $\rho(\underline{\beta})$, $\rho(\underline{1})$, and $\rho(\hat{\underline{\beta}})$ in the population for a given N , p , and Σ_{xy} . Point c represents $\epsilon\rho(\hat{\underline{\beta}})$ and point b represents a random value of $\rho(\hat{\underline{\beta}})$.

the density function of $\rho(\hat{\underline{\beta}})$ is presently unknown, an analytic solution is not now possible; and the use of empirical data presents numerous difficulties (Schmidt, 1970). On the other hand, the Monte Carlo approach allows the exact determination or very accurate estimation, for any given N , p , and Σ_{xy} of all of the values and distances of interest shown in Figure 1. $\rho(\underline{\beta})$ and $\rho(\underline{1})$ can be computed by means of the familiar formula for the correlation of a standardized variable with a composite:

$$\rho(\underline{W}) = \frac{W' \Sigma_y}{(W' \Sigma_x W)^{1/2}}, \quad (1)$$

where W is any given set of weights, here either $\underline{\beta}$ or $\underline{1}$. For $\rho(\hat{\underline{\beta}})$, for any given N and Σ_{xy} , a large number (v) of sets of $\hat{\underline{\beta}}_i$ can be computed and applied to the population, producing by equation (1) a distribution of $\rho(\hat{\underline{\beta}})$ comparable to the one in Figure 1. Then $\sum_{i=1}^v \rho(\hat{\underline{\beta}}_i)/v$ gives a very accurate estimate of $\epsilon\rho(\hat{\underline{\beta}})$. Not only v , but also N , p , and Σ_{xy} can be arbitrarily specified by the researcher. In addition the fact that the Monte Carlo approach works with the multivariate normal population means that the entire study is based on the random or correlational model of prediction, which is more appropriate than the fixed predictor model for most behavior science data (Burket, 1964). The value $\epsilon\rho(\hat{\underline{\beta}}) - \rho(\underline{1})$ is the mean superiority

(if any) of sample least square weights over unit weight for a given N , p , and Σ_{xy} . This is the value with which this study is primarily concerned.

The simulation approach is not entirely without problems however. In order to insure that the Σ_{xy} used are similar to those actually existing in the data domain of interest, it seems appropriate to employ sample correlation matrices (R) from the data domain as estimates of the Σ_{xy} . Although the R are, for all practical purposes unbiased estimates of the Σ_{xy} , sampling error in the R will cause the variance of the validities and predictor intercorrelations to be larger, on the average in the R than in the Σ_{xy} . This increased variance has the effect of increasing the average efficiency of regression relative to unit weights, thus exaggerating somewhat the utility of regression weights. Another effect to be expected is an increase in the incidence of suppressor variables.

There is another potential problem. Significant departures of empirical data from the assumptions of linearity, normality and homogeneity of conditional variances—assumptions basic to the regression model—apparently occur in about 20 per cent of empirical data samples (Sevier, 1957; Tupes, 1964; Ghiselli, 1964; Guion, 1965) and probably occur in a certain (but smaller) percentage of the empirical populations. Since such departures never occur in simulated populations, the results from Monte Carlo data may overestimate the efficiency of regression weights relative to unit weights in an empirical population. Because of these two considerations, Monte Carlo estimates of $\epsilon\rho(\hat{\beta}) - \rho(1)$ should be interpreted as maximal estimates of difference in performance between the two weighting methods.

Method

The Program

Sample correlation matrices (R) computed from randomly drawn samples from a multivariate normal distribution are distributed as $W(N, p, \Sigma_{xy})$, the Wishart distribution. Using this distribution, for each given N , p , and Σ_{xy} combination, 100 R matrices were generated and 100 $\hat{\beta}_i$ were computed, which, in turn produced 100 $\rho(\hat{\beta}_i)$ coefficients by equation (1). These coefficients were plotted to yield a $\rho(\hat{\beta})$ distribution as shown in Figure 1 and averaged to give $\epsilon\rho(\hat{\beta})$.

The process of R -generation works by means of the Bartlett decomposition of the Wishart distribution (Bartlett, 1933; Kshirsagar, 1959; Wijsman, 1957), and is outlined in Appendix 1. This approach to sampling from $N(\underline{\mu}, \Sigma)$ has been discussed and employed by Browne (1968) and Herzberg (1969), both of whom have carried out tests of the simulated data showing that the required assumptions are met. In addition Herzberg (1969) showed that the results from simulated data were almost identical with the results from a large sample of empirical data. Values of N and p lying within the ranges most frequently encountered in practice were chosen. N s investigated were 25, 50, 75, 100, 150, 200, 500, and 1000. Values of p used were 2, 4, 6, 8, and 10. For each N -level within each Σ_{zy} , the difference $\epsilon\rho(\hat{\beta}) - \rho(\underline{1})$ was computed, with each $\epsilon\rho(\hat{\beta})$ being based on 100 values of $\rho(\hat{\beta})$. The program also performed other operations not relevant here.²

Sampling Plan for $\hat{\Sigma}_{zy}$ Matrices

As mentioned above, sample correlation matrices are probably the best available estimates of population correlation matrices. Accordingly, a plan was set up to obtain a random sample of R matrices from the data domain of applied differential psychology. Four journals—*Educational and Psychological Measurement (EPM)*, *Journal of Applied Psychology (JAP)*, *Journal of Educational Psychology (JEP)*, and *Personnel Psychology (PP)*—were selected as representing the general area, and the years 1959–1969 were selected for examination. For two of the journals—*EPM* and *JEP*—the odd years were sampled and for the other two, the even years. In both cases, all correlation matrices of dimensions 3×3 to 11×11 were recorded. In some cases, only parts of larger matrices were used. Correlation vectors containing negative or zero values were not used as validity vectors, and it was sometimes necessary to rearrange rows and columns in order to meet this condition. An attempt was made to keep all validities above .20—a value chosen as approximately the minimum that would ordinarily be used in practice. For each matrix size, a random sample of 10 matrices was drawn from the pool of recorded matrices of that size and used as

² Requests for print-outs of the program should be sent to the author, Department of Psychology, Michigan State University, East Lansing, Michigan.

estimates of the Σ_{xy} . For certain of the larger matrices, a sample of 10 could not be obtained using only the volumes designated in the sampling plan, and additional samples had to be taken from the previously unused volumes. Even so, only eight 11×11 matrices could be found and two had to be taken from another source (Wechsler, 1949, p. 10). Obviously, the sampling fraction was much larger for the large than for the small matrices.

Results

Table 1 presents the mean superiority across all matrices of regression over unit weights ($\epsilon[\epsilon\rho(\underline{\beta}) - \rho(\underline{1})]$) for all 40 combinations of N and p . Negative values indicate that unit weights are superior to regression weights. When $N = \infty$, $\epsilon\rho(\underline{\beta}) = \rho(\underline{\beta})$ and $\epsilon[\epsilon\rho(\underline{\beta}) - \rho(\underline{1})]$ becomes $\epsilon[\rho(\underline{\beta}) - \rho(\underline{1})]$, the average difference in the populations between the correlations produced by the (error-free) population regression weights and unit weights (distance $a-d$ in Figure 1 for a given Σ_{xy}). The last column of Table 1 presents the average values of $\epsilon[\epsilon\rho(\underline{\beta}) - \rho(\underline{1})]$ across p -values at each N level.

An examination of the β for each of the 50 Σ_{xy} matrices revealed that 31 had one or more suppressor variables. Since suppressor variables are rarely used in applied differential psychology, it is probably hazardous to generalize to this data domain from the

TABLE 1

Mean Superiority across All Matrices of Regression Weights over Unit Weights [$\epsilon[\epsilon\rho(\underline{\beta}) - \rho(\underline{1})]$]
for all Combinations of N and p *

N	p					Mean Across p
	2	4	6	8	10	
25	-.0133	-.0295	-.0321	-.0873	-.1269	-.0578
50	.0120	.0118	-.0121	-.0092	-.0406	.0028
75	.0175	.0226	.0350	.0161	-.0095	.0163
100	.0225	.0296	.0488	.0308	.0057	.0273
150	.0228	.0358	.0610	.0486	.0232	.0383
200	.0249	.0393	.0662	.0514	.0297	.0423
500	.0269	.0455	.0769	.0643	.0449	.0517
1000	.0277	.0471	.0792	.0684	.0499	.0545
∞	.0286	.0491	.0831	.0725	.0551	.0577
No. of Matrices	10	10	10	10	10	50

* For any p value, for each of the 10 matrices, at each N level, $\epsilon\rho(\underline{\beta})$ is the average across 100 computed values of $\rho(\underline{\beta})$. The difference $\epsilon\rho(\underline{\beta}) - \rho(\underline{1})$ is averaged across the 10 matrices in the sample to give $\epsilon[\epsilon\rho(\underline{\beta}) - \rho(\underline{1})]$. The values in the last column are row averages.

present matrix sample. Therefore, Table 2, showing values of $\epsilon[\epsilon\rho(\hat{\beta}) - \rho(I)]$ for only those Σ_{xy} matrices without suppressors, was computed. The bottom row in Table 2 indicates the number of matrices at each p level without suppressors.

The ratio $\{\epsilon[\epsilon\rho(\hat{\beta}) - \rho(I)]/\epsilon[\rho(\hat{\beta}) - \rho(I)]\}$ is the difference in efficiency between unit and regression weights as a proportion of the maximum possible difference and is graphed in Figure 2 against sample size for all p -values for the entire sample of 50 Σ_{xy} matrices. Negative values indicate the superiority of unit weights. The value of N at which this ratio is zero is the sample size at which regression and unit weights are equally effective for a given number of predictors. Table 3 presents these "critical sample size" estimates calculated for the entire sample of Σ_{xy} matrices and also for the matrices without suppressors. In this Table, the critical sample size values for $p = 6$ in the entire Σ_{xy} sample and $p = 4$ in the suppressorless sample have apparently been distorted by error in the Σ_{xy} samples. Rechecks of the computations showed no errors. The "critical sample size" for $p = 6$ for the whole Σ_{xy} sample can probably safely be assumed to be somewhere near the midpoint between 44 and 60. For the suppressorless Σ_{xy} , the "break-even" sample size for $p = 4$ probably lies near the midpoint between 40 and 105.

Discussion

It should be noted in Tables 1 and 2 that, because of the idiosyncracies of individual matrix samples, the values of $\epsilon[\epsilon\rho(\hat{\beta}) - \rho(I)]$

TABLE 2

Mean Superiority of Regression Weights over Unit Weights $[\epsilon[\epsilon\rho(\hat{\beta}) - \rho(I)]]$ for Those Matrices Without Suppressors

N	2	4	p 6	8	10	Mean Across p
25	-.0233	-.0669	-.0652	-.1234	-.1342	-.0826
50	.0041	-.0349	-.0190	-.0418	-.0503	-.0284
75	.0084	-.0168	-.0047	-.0197	-.0224	-.0110
100	.0124	-.0109	.0121	-.0095	-.0119	-.0016
150	.0133	-.0033	.0209	.0065	-.0041	.0067
200	.0148	.0007	.0253	.0105	.0009	.0104
500	.0174	.0068	.0333	.0241	.0090	.0181
1000	.0180	.0089	.0361	.0280	.0188	.0206
∞	.0187	.0111	.0390	.0321	.0142	.0230
No. of Matrices	8	4	4	2	1	19

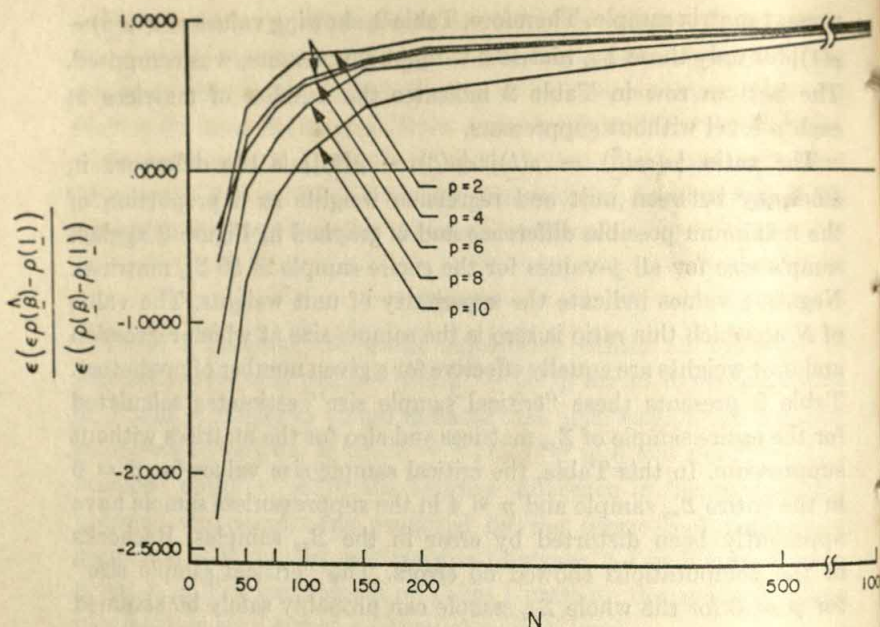


Figure 2. $\{\epsilon[\rho(\hat{\beta}) - \rho(\underline{1})]/\epsilon[\rho(\hat{\beta}) - \rho(\underline{1})]\}$ as a function of N for all values of p .

do not consistently decrease with increases in the number of predictors. With large samples of Σ_{xy} matrices, this would probably not be the case, but with samples of 10 matrices or less at each p value, the additional error introduced into the $\hat{\beta}_i$ as a result of the loss of one degree of freedom from the addition of one predictor is apparently not as important in determining differences between columns in Tables 1 and 2 as chance fluctuations, from matrix sample to matrix sample, in the value $\epsilon[\rho(\hat{\beta}) - \rho(\underline{1})]$. In Table 1, it can be seen that, for the entire sample of 50 Σ_{xy} matrices, unit weights are superior

TABLE 3
Critical Sample Size estimates for All Matrices and for Matrices Without Suppressors

p	All Matrices	No. of Matrices	Matrices Without Suppressors	No. of Matrices
2	37	10	40	8
4	44	10	256	4
6	44	10	105	4
8	60	10	142	2
10	92	10	194	1

to regression weights when the sample size is 25, regardless how few predictors are used. Even when only two predictors are used, a researcher can expect simple unit weights to be .0133 correlation points superior, on the average, to regression weights. When 10 predictors are used, the expected advantage for unit weights is .1269, a difference large enough to be of practical significance in many applied settings. The average superiority of unit weights over all p -values for sample size of 25 is .0578. Although the inferiority-superiority relationship is generally reversed when $N = 50$, the advantage of the regression weights is extremely small or even negative. Over all p -values this advantage is .0028.

The elimination of Σ_{xy} matrices with suppressors greatly reduces the matrix sample sizes upon which the values in Table 2 are based, but, because of the fact that suppressor variables are rarely used in applied differential psychology (e.g., see Adkins, 1946), these data are more relevant to the purposes of this study than the data in Table 1. In the suppressorless matrices, the relative advantage of regression weights—both sample and population weights—is less than it is in the entire sample of Σ_{xy} matrices. At a sample size of 25, the mean superiority of unit weights across p -value is .0826, quite a respectable difference. When $N = 50$, unit weights are superior by .0284 correlation units, on the average. When $N = 100$, unit weights are still superior in three out of the five p -values and as an average across all p -values.

For the entire sample of 50 Σ_{xy} matrices, across all levels of p , the sample size below which the applied psychologist can expect to suffer a decrease in the size of his obtained correlation as a result of using regression instead of unit weights is about 46. For only the Σ_{xy} matrices without suppressors, this value is 85. If we arbitrarily assume that .0150 correlation units is the minimum increase in predictive power, for most practical purposes, that will render the computation of regression weights worthwhile, then for the entire Σ_{xy} sample, averaging across levels of p , the minimum sample size needed is 60. For the matrices without suppressors, this minimum average sample size is 184. The implication of these latter figures is that, when applied psychologists are not actually suffering a loss in predictive power as a result of using regression weights, they are very often employing this complex statistical technique when it is probably a waste of time and effort to use it.

The "critical" or "break-even" sample sizes presented in Table 3 by number of predictors provide more detailed information about potential losses in predictive power resulting from the use of regression weights. Assuming that suppressors are not to be used, if an applied psychologist has only two predictors, he needs on the average, a sample of only 40 to insure no loss of predictive power from the use of regression techniques. If he uses six predictors, this figure jumps to 105. And if 10 predictors are used, he needs a sample of about 194. And it should be remembered that, for reasons discussed earlier, these "critical sample sizes" are best considered underestimates. Actual sample sizes needed to insure equality of performance of regression and unit weights are probably somewhat larger. It is not difficult to find in the differential psychology literature studies employing regression weights in which six or more predictors are used and the total sample size is below 105.

This finding of potential loss is apparently new. Many of the previous studies in this area raised the question of whether regression weights were really more efficient than simpler weighting methods, but none investigated, or even suggested, the possibility that the use of regression weights could result in a reduction in the size of obtained correlation.

Since many studies employing regression weights reported in the literature are characterized by sample sizes below the critical values presented in Table 3, it is concluded that many psychologists in applied areas are routinely penalizing themselves by their adherence to this statistical technique. The suggestion by Lawshe and Shucker (1959) that many psychologists are blinded to the possibilities of error in regression weights computed on small samples by the "apparent mathematical precision" of regression techniques appears to be borne out.

APPENDIX A

The Mathematics of The Monte Carlo Method: The Bartlett Decomposition Of The Wishart Distribution.

(1) Let $\Sigma_{xy} = XX'$

Σ_{xy} can be factored into XX' in a number of ways, all of which will work in this problem. In the program developed for this study, the factoring of Σ_{xy} was approached via the roots and vectors method.

(2) Let A be a $(p+1) \times (p+1)$ matrix defined as

$$A = 1/nTT',$$

Where T = a $(p+1) \times (p+1)$ lower triangular matrix, whose lower triangular elements are independent random variables:

$$T_{ii} \ (i > j) \text{ are } N(0, 1)$$

$$T_{ii} \text{ are } \chi(N-i)$$

$$T_{ii} \ (i < j) = 0$$

A is a sample variance-covariance matrix from a population where

$$\Sigma_{xx} = I, \text{ and } \chi(N-i) = \sqrt{\chi^2(N-i)}$$

(3) Let $C = N X A X' = X T T' X'$

Then $S = 1/NC$

S is a sample variance-covariance matrix from N (μ , Σ_{xy}) and is distributed as $W(N, p, \Sigma_{xy})$.

(4) C can be converted into R :

$$R = (\text{Diag } [C])^{-1/2} C (\text{Diag } [C])^{-1/2}$$

Then R is distributed as the maximum likelihood estimate of Σ_{xy} based on samples of N observations from a $p+1$ normal distribution with a population correlation matrix, Σ_{xy} (Browne, 1968).

(5) Generation of the T matrix. There are a number of techniques for generating pseudorandom numbers from a rectangular distribution and transforming these into random normal deviates and random numbers from Chi distribution required for the T matrix (e.g., see: Hull and Dobell, 1962; Tausky and Todd, 1956; Muller, 1959; Teichroew and Sitgraves, 1961). In this study, the pseudo-random numbers were taken from a subroutine in the core of the 6500 CDC computer, called $RANF(X)$.

(a) These pseudo-random numbers, w_i , are transformed into pseudorandom normal deviates, g_k , by the expressions (Box and Muller, 1958):

$$g_k = (-2 \log_e w_i)^{1/2} \cos 2\pi w_{i+1}$$

$$g_{k+1} = (-2 \log_e w_i)^{1/2} \sin 2\pi w_{i+1}$$

(b) Pseudo-random numbers $\chi_i(f)$ from a Chi distribution with f degree of freedom are obtained from:

$$\chi_{i(2f)} = \left(-2 \sum_{i=1}^f \log_e w_i \right)^{1/2}$$

$$\chi_{i(2f+1)} = \left(-2 \sum_{i=1}^f \log_e w_i + \vartheta_i^2 \right)^{1/2}$$

(Box and Muller, 1958; Kendall, II, 1946, pp. 132-133).

- (c) Once generated, random normal numbers can be filled in at random in the lower triangular cells of T . Random elements from the Chi distribution are filled in on the diagonal of T , but only in the order of the degrees of freedom of the distributions from which they are drawn [See (2)].
- (d) Note: (1) For any p , N , and Σ_{xx} , a different T matrix must be generated for each R matrix sampled, and thus for each $\rho(\beta)$ computed.
- (2) The program was set up to print out the last random number from the $RANF(X)$ function at the end of each run. This number was then punched into a card and used to re-enter the random number sequence at the start of the next run. Thus each R was generated from a different random number sequence.

REFERENCES

- Adkins, D. C. *Construction and analysis of achievement tests*. U. S. Government Printing Office, 1947.
- Anderson, T. W. *Introduction to multivariate statistical analysis*. New York: Wiley, 1958.
- Bartlett, M. S. On the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh*, 1933, 53, 260-283.
- Box, G. E. P. and Muller, M. E. A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 1958, 29, 610-611.
- Brogden, Hubert E. On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology*, 1946, 37, 64-76.
- Browne, Michael W. A comparison of factor analytic techniques. *Psychometrika*, 1968, 33, 267-334.
- Burket, George R. A study of reduced rank models for multiple prediction. *Psychometric Monograph*, No. 12, 1964.
- Cureton, Edward E. Validity, reliability and baloney. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1950, 10, 94-96.
- Ewen, R. B. Weighting components of job satisfaction. *Journal of Applied Psychology*, 40, 1956, 73-77.
- Ghiselli, E. E. Dr. Ghiselli comments on Dr. Tupes' note. *Personnel Psychology*, 1964, 17, 61-63.

- Grant, D. L. and Bray, D. W. Validation of employment test for telephone company installation and repair occupations. *Journal of Applied Psychology*, 1970, 54, 7-15.
- Guilford, J. P., Lovell, C., and Williams, R. M. Completely weighted versus unweighted scoring in achievement examination. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1942, 2, 15-21.
- Guion, R. M. *Personnel testing*. New York, McGraw-Hill, 1965.
- Harper, B. P. and Dunlap, J. W. Derivation and application of a unit scoring system for the Strong Vocational Interest Blank for Women. *Psychometrika*, 1942, 7, 289-295.
- Herzberg, Paul A. The parameters of cross-validation. *Psychometric Monograph*, No. 16, 1969.
- Hull, T. E. and Dobell, A. R. Random number generators. *Society for Industrial and Applied Mathematical Review*, 1962, 4, 230-254.
- Kendall, M. G. *The advanced theory of statistics*, Vol. II. London: Charles Griffin, 1946.
- Kirkpatrick, James J. Cross-validation of a forced-choice personality inventory. *Journal of Applied Psychology*, 1951, 35, 413-416.
- Kshirsagar, A. M. Bartlett decomposition and Wishart distribution. *The Annals of Mathematical Statistics*, 1959, 30, 239-241.
- Kurtz, Albert K. A research test of the Rorschach test. *Personnel Psychology*, 1948, 1, 41-51.
- Lawshe, C. H. and Schucker, R. E. The relative efficiency of four test weighting methods in multiple prediction. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1959, 19, 103-14.
- Muller, M. E. A comparison of methods for generating normal deviates on digital computers. *Journal of the Association for Computing Machinery*, 1959, 6, 376-383.
- Phillips, A. J. Further evidence regarding weighted versus unweighted scoring of examinations. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1943, 3, 151-155.
- Ryans, David G. An analysis and comparison of certain techniques for weighting criterion data. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1954, 14, 449-457.
- Schmidt, F. L. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. PhD dissertation, Purdue University, 1970.
- Sevier, Francis A. C. Testing the assumptions underlying multiple regression. *Journal of Experimental Education*, 1957, 25, 323-330.
- Tausky, O. and Todd, J. Generation and testing of pseudo-random numbers. In H. A. Meyer (Ed.). *Symposium on Monte Carlo methods*. New York: Wiley, 1956, pp. 15-28.
- Teichroew, D. and Sitgraves, R. Computation of an empirical sampling distribution for the *W* classification statistic. In H. Solomon (Ed.), *Studies in item analysis and prediction*. Stanford: Stanford University Press, 1961, pp. 252-275.

- Trattner, M. H. Comparison of three methods for assembling aptitude test batteries. *Personnel Psychology*, 1963, 16, 221-232.
- Tupes, E. C. A note on "validity and nonlinear heteroscedastic models." *Personnel Psychology*, 1964, 17, 59-61.
- Wechsler, D. *The Wechsler Intelligence Scale for Children*. New York: Psychological Corporation, 1949.
- Wesman, A. G. and Bennett, G. K. Multiple regression vs. simple addition of scores in prediction of college grades. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1959, 19, 243-246.
- Wijisman, R. A. Random orthogonal transformations. *The Annals of Mathematical Statistics*, 1957, 28, 415-423.

TRUE SCORE THEORY: A PARADOX

J. O. RAMSAY

McGill University

CLASSICAL mental test theory, as set down by Guliksen (1950), was thought to depend fundamentally on four assumptions: (a) The mean of error is zero, (b) the correlation between true score and error is zero, (c) the correlation between errors on two occasions with the same test is zero, and (d) the correlation between true score on one occasion and error on another is zero. These assumptions are held to be true in the population of test scores and individuals. More recently, however, Novick (1966) has shown that the last three assumptions follow from the first and the assumption of linear independence of error over testing occasions for a particular individual.

Although this work has clarified greatly the foundations of mental test theory, it is now clear that there are few assumptions more ubiquitous in any body of theory than the notion that expected observed score is equal to true score. Lord and Novick (1968) discuss this relationship in some detail and reach the conclusion that if there is no a priori reason for accepting this statement (i.e., no platonic true score), then it is usually not unreasonable to simply define true score as the expected value of observed score. It is important to ask whether there are consequences of this assumption that run against intuition or established tradition. This paper will attempt to show that there are such consequences.

Test reliability is defined in one of two ways. If it is defined as the correlation between test scores on one occasion with those on another completely independent occasion, then a direct consequence is that it is the ratio of true score variance to observed score variance, or:

$$\rho = \frac{\sigma_t^2}{\sigma_o^2} \quad (1)$$

Alternatively, it may be defined directly as (1), which may also be expressed as

$$\rho = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2} \quad (2)$$

where σ_e^2 is error score variance.

Since all test scores are distributed on a closed interval in practice, and usually in theory as well, it suffices to consider the behavior of reliability for scores distributed on the interval $[0, 1]$. Test scores of this kind can have zero reliability only when true score variance is zero. That is, error score variance must be finitely large. To make this more clear, σ_e^2 can be expressed in the following form:

$$\begin{aligned} \sigma_o^2 &= \sigma_t^2 + \sigma_e^2 \\ &= E[\text{Var}(x | t)] + \text{Var}[E(x | t)] - \sigma_t^2 \\ &= E[\text{Var}(x | t)] \end{aligned} \quad (3)$$

by using a fundamental theorem about variance and noting that $E(x | t) = t$. Now the variance of observed score for a particular true score is very definitely limited. Moreover, the constraint that the mean of this distribution must equal the true score bounds this variance by that of a Bernoulli variable with mean equal to the true score. That is,

$$\text{Var}(x | t) \leq t(1 - t). \quad (4)$$

In order to see just how different from zero this lower limit on reliability is apt to be in practice, it is necessary to propose some more or less realistic models for the distribution of true score itself. The logical candidate seems to be the beta distribution which can be made to look like the uniform or bernoulli but also which can reproduce the kind of unimodal distribution of scores that one expects in practice. Therefore, let

$$f(x | t) = \frac{1}{B(a + 1, b + 1)} x^a (1 - x)^b \quad (5)$$

and

$$g(t) = \frac{1}{B(c+1, d+1)} t^c (1-t)^d. \quad (6)$$

The parameters, c and d , in (6) are chosen to give $g(t)$ a mean of μ_t and a variance of σ_t^2 by using

$$d = \frac{\mu_t(1-\mu_t)^2}{\sigma_t^2} + \mu_t - 2 \quad (7)$$

$$c = \frac{1 - \mu_t(d+2)}{\mu_t - 1} \quad (8)$$

The parameters, a and b , are chosen to provide $f(x|t)$ with a mean of t and the maximum possible variance by using (7) and (8) and the fact that the maximum variance when $a \geq 0$ and $b \geq 0$ is given by

$$\sigma_{x|t}^2 \leq \begin{cases} \frac{t^2(1-t)}{1+t}, & 0 \leq t \leq .5 \\ \frac{t(1-t)^2}{2-t}, & .5 \leq t \leq 1.0. \end{cases} \quad (9)$$

The expression for the expected observed score variance is then

$$E[\text{Var}(x|t)] = \int_0^1 \int_0^1 (x-t)^2 f(x|t) g(t) dx dt \\ = \int_0^1 \left[\frac{(a+1)(a+2)}{(a+b+2)(a+b+3)} - \frac{2t(a+1)}{(a+b+2)} + t^2 \right] g(t) dt. \quad (10)$$

If this is evaluated numerically and substituted along with σ_t^2 into (2), the result is a "realistic" lower bound on reliability as a function of true score mean and variance. It is not, of course, the absolute lower bound on reliability, but it is a lower bound consistent with what intuition demands of the distribution of true score and consistent with the assumption that expected observed score is equal to true score.

Figure 1 shows this lower bound as a function of true score variance for a mean true score of 0.6. Translated into classroom testing language, it says that if the average student's true score is 60 per cent and the standard deviation of true scores is 15 per cent, then the test cannot be less reliable than 0.28. If the standard deviation of true scores is 20 per cent, then test has a reliability of at least 0.45.

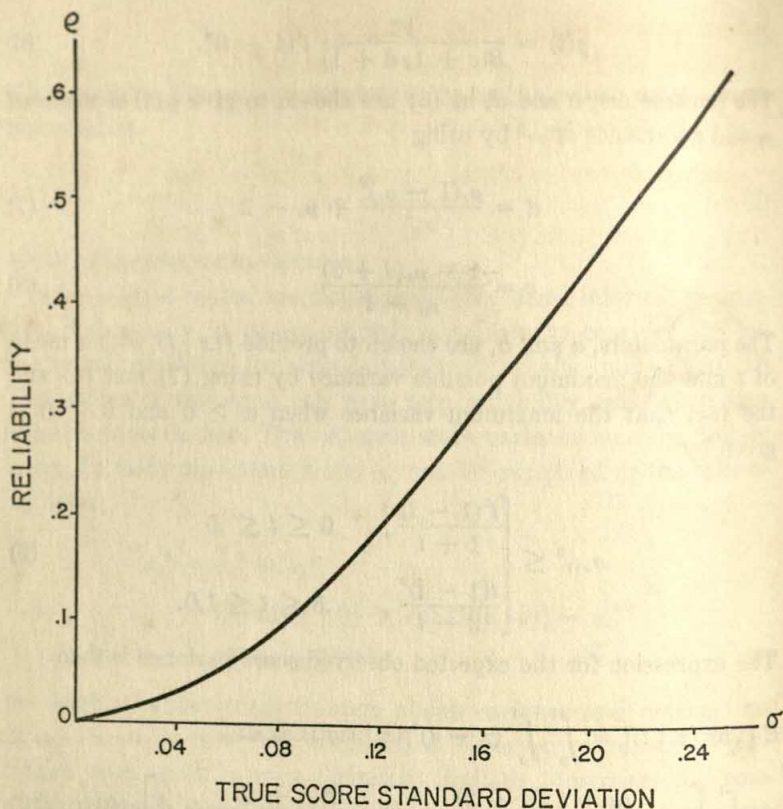
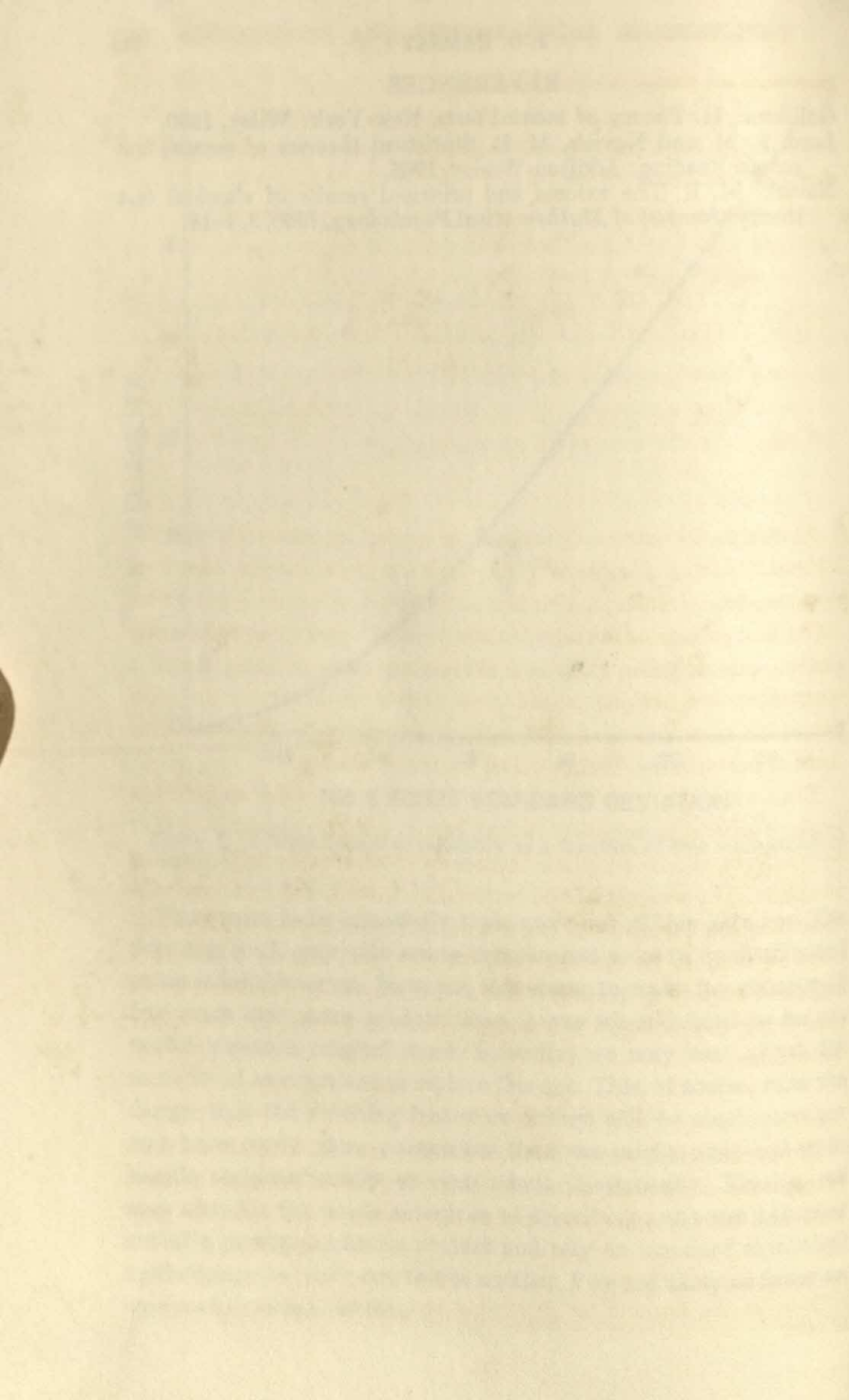


Figure 1. A lower bound on reliability as a function of true score standard deviation.

There seem to be essentially three ways out of this paradox. The first is to work only with scores transformed so as to be distributed on an infinite interval. However, this seems to make the concept of true score even more artificial than it was when defined to be expected observed original score. Secondly, we may cast about for some set of assumptions to replace this one. This, of course, runs the danger that the resulting test score theory will be much stronger and have many more parameters than we might wish either to handle computationally or even admit theoretically. Finally, we may abandon the whole enterprise of describing test score behavior out of a purely predictive context and rely on standard statistical methodology to relate one test to another. Few are likely to favor an approach as radical as this.

REFERENCES

- Gulliksen, H. *Theory of mental tests*. New York: Wiley, 1950.
- Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores*. Reading: Addison-Wesley, 1968.
- Novick, M. R. The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 1966, 3, 1-18.



A TEST OF THE TRAIT-VIEW THEORY OF DISTORTION IN MEASUREMENT OF PERSONALITY BY QUESTIONNAIRE

SAMUEL E. KRUG¹ AND RAYMOND B. CATTELL

University of Illinois

ATTEMPTS to correct questionnaire scores for distortion and instrument factor ("content") effects have been largely based on response set (Cronbach, 1946), multi-method (Campbell and Fiske, 1959) and social desirability (Edwards, 1957) conceptions. Recently Cattell and Digman proposed subsuming most of these under a comprehensive *perturbation theory* (1964) with corrections indicated for questionnaires and ratings according to the specialized form of perturbation theory called trait-view theory.

Trait-view theory proposes to consider the distortion as itself the product of the personality factors one is out to measure, together with effects of role adoption tendencies more specific to the testing situation. The weights of the personality factors and dynamic role-identification factors will vary with different motivational situations in testing. By finding their weights, and storing them for a sufficient variety of testing situations, it should be possible for the practicing psychologist routinely to correct scores for the above distortions.

Description of Trait-View Theory

In psychological terms the theory states that an individual's misperceptions of himself on a particular trait is a function of that trait and all his other traits, plus motivation specific to the par-

¹ Now at the Institute for Personality and Ability Testing, Champaign, Illinois.

ticular role. For example, a person high on super-ego strength (G on the 16 P.F.) might, in the first place, understate his own excellence on that trait, and an intelligent person might more clearly recognize the dominance motivations in his own behavior and thus overscore, relative to an individual who is less intelligent, on a dominance scale. Again, in responding to a question which indicates his degree of unreleased tension, a person with a very highly developed self-sentiment might be unable to recognize the full extent of his tension, since it conflicts with his self-concept. He might, therefore, give an estimate of his behavior which is well below the true situation.

Stated as a model, we have, in specification equation form:

$$S_{i,c} = b_{i,c}T_{ic} + \cdots + b_{i,i}T_{ii} + \cdots b_{i,k}T_{ki} + b_{i,c}T_{i,c}$$

where T_{ic} represents the individual's score on the true factor c (as distinct from his obtained score, S_{jet} , on any trait scale) and the several T_{jet} terms represent the individual's true scores on other traits which are involved with the performance on S_{jet} .

It is clear that what must be done in order to obtain a true measurement on the factor is to solve for the unknowns in this equation and thus provide a means both of estimating the true factor and of establishing which particular personality factors enter into the distortion between T_{ic} and S_{jet} .

It is apparent that if an individual is given a psychological test as part of a job selection procedure, the greater his interest in obtaining the job (and the greater his knowledge of what is required), the more his answers will tend to come in line with his perception of the required 'type.' Similarly, a mother who is required to complete a questionnaire or answer some questions which are designed to help a school counselor advise her child more accurately, may tend to give completely accurate reports or somewhat jaded reports depending upon how 'motherly' she is feeling at the time. The final result in each case is a personality profile which, for one reason or another, is distorted.

Role perception almost constantly has a powerful distortion effect in personality questionnaires. In an analysis of the concept and measurement of role behavior, as distinct from personality, Cattell (1963) has proposed that role be recognized as an additional factor beyond the usual personality factors. It differs from a general

personality factor in that it enters only a relatively specific subset of situations, whereas the former shows across all roles.

Using the role theory requires us at this point to adopt two suppositions, neither of which has as yet actually been investigated. The first says, as above, that we should add a measure of role behavior in the statistical form of an ordinary primary personality factor—actually a dynamic factor and a relatively specialized kind that can be called a role factor, R . It is possessed by those who have experience in the role to varying degrees and not by others. The second supposition says that unlike a personality trait it is 'modulated' in level by the situation itself, that is, the situation can stimulate it to new levels. We may add thirdly, as a practical test proposition, that it is especially desirable to measure this role factor by objective (decidedly less distortable) motivation measurement devices (Cattell, Radcliffe, and Sweeney, 1963).

In this first investigation we do not propose to separate the contributions from (a) the role strength and (b) its modulation in the given situation. It suffices for the aspect of trait view theory here studied if we measure the total effect of role involvement as represented by T in the equation:

$$S_{iit} = b_{iit}T_{ic} + \cdots + b_{iit}T_{is} + \cdots + b_{iit}T_{kt} + b_{iRi}R_{ie}$$

Design of the Experiment

The aim of the present research is to test the hypotheses: (1) that the score of an individual on a scale set up to measure a known source trait will be significantly determined not only by his real level on that source trait, but also by the influence of other personality factors in a characteristic and meaningful distortion of the self-evaluation in the test-taking behavior, and (2) that the weights of the personality factors influencing the score will differ according to the definition of the test-taking situation, and that when this situation includes instructions which give rise to the adoption of a role, a distinct role factor will also emerge, influencing the total score. As a third question, but in the applied field, it is asked whether the true score can be better estimated in a given situation by employing (a) a measure of the extent to which the role factor operates for each individual, and (b) a measure of the role factor together with weighted scores on the other personality factors.

It will be asked at the same time whether the estimation of the pure source trait factor is different from that obtained simply on the usual scale for the source trait itself.

The design for the experiment naturally required that the measures for a sufficient number of source traits be employed both in order that the factor equation be soluble and in order to deal with a really substantial part of the total personality effect upon any single scale.

For this purpose, the Sixteen Personality Factor Questionnaire—16 P.F.—(Cattell, Eber, and Tatsuoka, 1969) was selected, since it samples sixteen dimensions of the personality sphere. Moreover, in order to insure greater reliability, two forms of the test, A and B, were employed.

The subjects were 159 first- and second-year undergraduates from four colleges in the United States.

The design of the study further required that the same trait scales be administered under different conditions or role-involvement situations, four in all. The four test-taking role situations were selected in part for their theoretical interest and in part for their practical importance and their adaptability to an experimental setting. In each situation, the subjects were given both forms of the 16 P.F. and a short, objective motivation measurement scale which was constructed to assess their degree of role-involvement in the particular situation.

The design of this role-involvement measure utilizes the findings of a long series of experiments on objective motivation measurement. Approximately half of the items had already been used in the Motivation Analysis Test (Cattell, Horn, Radcliffe, and Sweney, 1964) and were of proven validity against the career and other factors therein. The rest were constructed to meet the particular needs of the present study.

The following four situations were included in the present study:

(1) *The standard cooperative research situation.* Since many investigations are done under conditions where the subject is simply sitting for an experiment, is guaranteed anonymity, and is assured absence of any consequences from his performance, this is an important standard situation to be evaluated.

(2) *The job seeking situation.* In this situation, the subjects were instructed to fill out the questionnaire as part of a study to

evaluate their potential for a future job in teaching. More than half of the subjects were education majors, so this was not merely "acting" a role. They were told that the results would be used by the college of education to determine the effectiveness of their training. In this case, the individual is filling out the questionnaire in such a way that another individual or institution is evaluating him and it is reasonable to assume that such a situation will introduce an element of 'desirability' faking or 'stereotype' faking. Although the particular type of job selected may dictate specific aspects of distortion, nevertheless, one may assume that some general elements will be sampled which may allow certain generalization to the job seeking situation in general.

(3) *The ideal self-distortion.* This situation was selected partly because of its theoretical interest for general research on the self-concept and partly to check on the implied assertion above that the job seeking situation has many specific elements and is not identical to a situation in which the individual gives the best profile he can of himself. In other words, there should be some difference between the individual's *generalized* ideal, and his estimate of the ideal required for a particular situation.

(4) *The 'Operation Match' or Prospective Dating situation.* In this situation, the subjects were asked to fill out the questionnaire for a study being done by Operation Match, a computer dating service. Again, this situation introduces desirability faking of a type—that involved in appearing attractive to the opposite sex—but it is assumed that this would be somewhat different from that encountered in either (2) or (3) above.

It should be recognized incidentally, that even the standard situation (1) is not conceived to be free from distortion in terms of the theory which is being employed here.

Analysis of Data

From each subject, 72 scores were obtained, 16 trait scores (Form A and Form B scores combined) and two role scores on each of four occasions. Principal axes factors were obtained from the unreduced correlation matrix. The roots were plotted to determine the number of factors to be retained for further analysis, as indicated by the Scree test (Cattell, 1966). Twenty-one factors were indicated and retained. Convergent communalities were estimated for 21

factors by traditional iterative principal axes procedures. The orthogonal, principal axes solution was rotated to the first approximation of the oblique, simple structure position by a Procrustes procedure (Hurley and Cattell, 1962).

Following this, graphical rotations were performed to approach the simple structure position with somewhat greater precision. Finally, the Maxplane program (Cattell and Muerle, 1960) was used to "clean-up" the structure. The ultimate hyperplane count was 76.3 per cent, which is high among published researches and indicates a sufficiently stable and, therefore, psychologically meaningful position.²

Examination of the factor patterns for the 16 scales shows the presence of those differences in the contributions of various personality factors to particular scales that the theory would require. However, due to present disputes about the evaluation of the significance of the difference between two factor loadings it is not easy to give an acceptable appraisal of the significance. If, using the same principle as in Harris's (1965) test for the significance of a loading we consider these as partial correlations and calculate approximately, considering the average inter-factor r to be the general interfactor r , a loading significant at the .05 level would be approximately .16. By the same principle differences of .10, .08, and .06, with the lower r at .05, .10, and .15 respectively would be significant at the .05 level.

Psychologically, it is interesting to note that among the suggestive differences are: (a) a tendency for the outgoing warmth of factor A to favor higher estimation of factor O (worrying, guilt-prone) more strongly in the dating than in the job-seeking situation, (b) a tendency for the dominance (E) factor to favor higher estimation of factor H (venturesome) in the job-seeking than in the anonymous situation, and (c) a tendency for high ergic tension level (Q_4) to favor higher estimation of factor O in the job-seeking than in the self-sentiment situation. It is not difficult to generate psychological theories for these, from the understanding of the general nature of the personality source traits and these situations.

² The complete factor pattern matrix has been deposited with the National Auxiliary Publications Service, CCM Information Corp., 909 Third Avenue, New York, N. Y. 10022. Order Document No. 01510. Remit in advance \$5.00 for photocopies or \$2.00 for microfiche.

In addition to this new analysis according to trait-view theory, by factor analysis, we made the more traditional comparisons to determine whether there were significant differences in mean scores on each factor over the different role conditions. The results of separate *t*-tests for each variable between raw scores in each condition and the mean of all four conditions, which may be thought of as an averaged role condition, are given in Table 1. While these tests may not be considered conclusive, because of the possibility of correlated error terms, they may at least be considered indicative of real shifts among the set of role conditions. The similarity to the shifts under the ordinary conscious "fake good" instruction may be seen from column 5 in Table 1, taken from the 16 P.F. Handbook (Cattell, et al., 1969).

In order to test the third hypothesis, that there is practical utility to the theory in that the true factor may be better estimated by using scores on role variables as well as scores on other personality variables, estimates of the true factor scores were next made in various ways, separately for each of the four test-taking conditions. First, the factor was "estimated" in the ordinary way, simply

TABLE 1

Differences in Personality Scores of Each Role Condition From the Mean Role Condition

Factor	Condition 1	Condition 2	Condition 3	Condition 4	Mean Role Condition	Under instructions to "fake good"*
A	5.7**	6.9*	7.3**	6.7	6.7	7.1
B	6.7	6.6	6.4	6.7	6.6	6.0
C	5.4**	7.3	8.3**	7.1	7.0	7.7
E	5.9	5.7	5.3*	5.9	5.7	5.6
F	6.2**	5.5**	6.1	6.2	6.1	6.5
G	5.2**	7.3**	7.7**	5.8**	6.5	6.9
H	5.5**	7.4*	7.8**	7.2	7.0	7.9
I	6.0	6.2*	5.8	5.6	5.9	4.9
L	5.6**	4.0	3.5**	4.1	4.3	4.2
M	6.3**	5.8	5.2**	5.6	5.7	4.9
N	5.3**	6.4	6.7**	6.1	6.1	6.8
O	5.9**	3.6**	2.9**	4.0	4.1	3.4
Q ₁	6.0**	6.9	7.0*	6.7	6.6	6.6
Q ₂	5.6**	4.8	4.3**	4.7	4.9	4.1
Q ₃	5.2**	7.7**	8.1**	6.6	6.9	7.7
Q ₄	5.7**	2.9**	2.1**	3.5	3.6	2.5

Note.—Scores are expressed here in terms of the sten scale, having a mean of 5.5 and a standard deviation of 2.

* Taken from the Handbook for the 16 PF (Cattell, et al., 1969).

** $p < .05$.

** $p < .01$.

from the scale which was designed to measure it. The second estimate was made by taking the regression of all the personality variables (only for that situation) on the particular true factor source trait.³ The third estimate was made by taking the regression of the appropriate personality variable only on the two role variables for that condition. Finally, the factor was estimated by taking the regression of all personality and both role variables, i.e., combining three and four described above.

In the first case this amounts simply to taking the correlation between the scale and the pure factor as calculated from the factor structure matrix. In the last three cases, the degree of validity is indicated by the multiple correlation coefficient between (a) personality variables, (b) role variables, and (c) both of these, with the pure factor. These values are shown in Table 2. Summed values from Table 2 represent actually the means when the correlations were transformed to Fisher Z coefficients and back to r 's. On these Z values, analyses of variance were carried out separately for each role condition to test whether there was a significant difference in factor estimation among the four approaches. A single factor design for correlated between group observations was used (Winer, 1962). After determining that the individual F ratios were significant, differences between condition means were examined separately by the method of Scheffe (cf. Hays, 1963). The results of the analyses of variance and the post-hoc comparisons are presented in Table 3.

Discussion and Interpretation

The first two hypotheses stated earlier, (1) that the score of an individual on a factor source trait will be determined not only by his actual score on the scale for that trait but also by other personality factors, and (2) that the weights on these will differ according to

³The reader should be reminded of the distinction, here and in Table 2 between the aim of obtaining a correction for a role as here, and the aim of what Eber and Cattell (1968) have called *computer synthesis scoring* or variance allocation. In the latter it is assumed that regardless of test situation, some scales contain displaced variance that should be returned to other scales. To do that one averages across all possible role situations. In the present method one uses specifically the weights for the known test taking situation. In the role correction calculation, a constant peculiar to the equation needs to be added to allow for the mean shift of *everyone* in going into a situation, as shown in Table 1.

TABLE 2

Validities of Source Trait Estimation in Four Test-Taking Role Situations

Source Trait	Situation 1—Anonymity Method				Situation 2—Job Seeking Method			
	1	2	3	4	1	2	3	4
A Affectothymia	75	77	76	79	74	77	77	79
B Intelligence	69	71	69	72	81	85	81	85
C Ego Strength	73	81	75	81	68	74	69	75
E Dominance	61	70	62	71	56	66	55	66
F Surgency	78	83	78	83	59	72	58	73
G Super Ego	58	59	57	58	80	83	81	84
H Parmia	41	51	45	51	49	53	51	57
I Premia	81	84	81	84	70	74	69	73
L Protension	70	73	71	73	67	75	67	75
M Autia	51	50	53	50	16	48	14	47
N Shrewdness	66	71	66	71	61	66	62	69
O Guilt	-08	36	10	39	07	56	24	61
Q ₁ Radicalism	62	70	66	71	68	71	67	70
Q ₂ Self-Sufficiency	56	62	56	64	68	73	69	73
Q ₃ Self-Sentiment	56	57	56	57	47	51	66	68
Q ₄ Ergic Tension	59	66	58	67	82	87	84	89
	62	69	64	69	63	71	65	73
Source Trait	Situation 3—Ideal Self Method				Situation 4—Dating Method			
	1	2	3	4	1	2	3	4
A Affectothymia	61	65	61	66	55	66	54	66
B Intelligence	81	84	82	85	87	88	87	88
C Ego Strength	65	71	65	71	73	77	73	77
E Dominance	47	62	48	64	58	61	62	64
F Surgery	43	50	44	50	64	69	64	69
G Super Ego	30	50	30	49	66	67	67	67
H Parmia	17	49	35	54	37	44	44	47
I Premia	65	71	64	72	83	86	83	86
L Protension	53	57	56	57	77	81	77	81
M Autia	67	73	67	73	51	56	50	57
N Shrewdness	54	65	54	65	56	68	59	68
O Guilt	48	62	50	63	02	50	10	50
Q ₁ Radicalism	64	71	64	71	73	78	74	78
Q ₂ Self-Sufficiency	68	74	69	73	63	68	63	67
Q ₃ Self-Sentiment	66	70	66	71	38	46	37	47
Q ₄ Ergic Tension	39	48	37	47	78	79	79	80
	56	65	57	66	64	70	65	71

Key to Methods: 1 = Ordinary scale scoring; 2 = Using all personality factors; 3 = Using dynamic role factors; 4 = Combining 2 and 3.

Note.—All multiple correlations have been corrected for shrinkage.

TABLE 3

Mean Differences in Factor Estimation Procedures: Results of Analyses of Variance and Post-hoc Comparisons

Condition 1- $F = 21.00^{**}$			
	Personality	Role	Personality + Role
V_{fs}	+.07**	+.02**	+.07**
Personality		-.05**	.00
Role			+.05**
Condition 2- $F = 25.00^{**}$			
	Personality	Role	Personality + Role
V_{fs}	+.08**	+.02*	+.10**
Personality		-.06**	+.02
Role			+.08**
Condition 3- $F = 50.00^{**}$			
	Personality	Role	Personality + Role
V_{fs}	+.09**	+.01	+.10**
Personality		-.08**	+.01
Role			+.09**
Condition 4- $F = 26.67^{**}$			
	Personality	Role	Personality + Role
V_{fs}	+.06**	+.01	+.07**
Personality		-.05**	+.01
			+.06**

Note.—The values in this table represent the algebraic differences in mean validity between the method designated by the column heading and the method designated by the row heading. Thus, under condition one, .07 is the difference of the method using the regression of all personality factors in estimating the pure factor and the ordinary single scale (V_{fs}) approach, the former being more valid than the latter.

* $p < .05$.

** $p < .01$.

the test taking situation and will include contributions specifically from a role factor, can be considered together.

The source traits operation in the 16 P.F. can, on the whole, be readily identified by the marker variable loadings. Factors 1 through 16 were hypothesized to be the source traits measured by the 16 P.F. With the exception of factors 7 and 12 (source traits H and O) all have their highest loadings on the appropriate marker variables. Factor 7, while not having its highest loading on the marker variables, does have reasonably high values. In addition, other variables which are highly loaded by this factor, such as E in the general experimental role, are known from previous researches (Cattell, 1957) to show these moderately high loadings. Such inter-correlations among primaries in fact provide the basis for the second-stratum pattern of anxiety, such as C, L, and H.

The patterns expected for the role factors are somewhat less clear.

Factor 18 is fairly obviously a career role factor as indicated by its marker variables. Factor 20 can be interpreted as the role factor for the Operation Match (Dating) situation from its loading of .81 on the first role variable for that condition. Although there is a higher loading of .83 on variable E in the Dating role factor, we might reasonably hypothesize that in males, at least, dominant, assertive behavior is part of the mate-finding pattern.

Factors 17, 19, and 21 still remain unidentified in the factor matrix. The means of the absolute values of the loadings of factor 17 for each of the four situations are .276 for the general condition, .086 for the career situation, .121 for the self-sentiment distortion, and .131 for the Operation Match condition. The relative strength of its appearance in the variables measured in the general condition suggests a tentative identification of factor 17 as a general experimental role factor. For factor 19, the same mean values are .111, .186, .209, and .183. Noting that the highest loading of 19 is in the ideal self-situation and other high loadings are on G(+), E(-), and C(+), we tentatively consider 19 a self-sentiment role factor, i.e., a strength of interest in the ideal self.

A factor such as 21 was originally hypothesized to appear purely as an instrument factor which would have positive loadings on the role variables (because they are measured by *objective* motivation devices) and negative or zero loadings on the personality variables (measured by questionnaire). In the final rotation it would only roughly fit the hypothesized pattern.

In testing the second hypothesis: that significant and significantly different weights will be found on other factors than the source trait being measured, as the test-taking situation changes, it is necessary to bear in mind that the search for simple structure will—as far as errors occur in rotation—operate in the direction of reducing the incidence of such values. However, it cannot do so beyond a certain point, for the nature of simple structure is such that reducing some would tend to raise others. We should recognize also that in estimating a pure trait, we are concerned with the V_{fe} , not the V_{fp} . In other words the weights of the various scale scores in estimating a true source trait are what we need to compare from situation to situation, from the practical test point of view, and these will respond to changing correlations of the traits from situation to situation too. Whether this latter kind of change also occurs

is a matter for later and larger researches to settle. A rough idea of the amount of role involvement may be gleaned from the fact that the average number of significant loadings per variable is four (not one) in the role taking situations.

That the contributions of the source traits to determination of scale score change from role to role was mentioned above in examining successive rows of the factor pattern matrix. The contributions of source traits E and G to scale A change direction from the first condition to the second condition. This change in direction of loadings was evident throughout the matrix. The change in value of the marker loadings is also quite interesting. The variance of scale H (venturesome) in condition two seemed to be taken up primarily by factor E (assertiveness) and factor O (confidence), which makes a good deal of sense in the job seeking situation. Other examples are factor E, which seems to operate more forcefully on scale E in the career setting than in the self-sentiment distortion, or factor F (enthusiastic) which operates more strongly in the career situation than in the Operation Match situation. In an attempt to find how significant these changes are, we used, in addition to the pair-wise comparison of loadings mentioned above, the Cochran Q statistic. This evaluates whether the true proportion of salient loadings was constant across all role conditions for each variable. Only one variable, Q_2 (group dependence), showed a significant shift ($Q = 18$; $p < .001$) in the number of salient loadings across the various conditions, but this does not negate the pair-wise findings.

The final hypothesis to be examined is that a more valid estimate of any source trait factor can be made by taking the regression of other personality and motivation scales (along with the appropriate personality scale) than can be done from the single scale itself. Table 3 above shows the outcome of this investigation. Four methods are actually compared here: (1) the simple scale, (2) using other personality scales, (3) using role factors, and (4) using (2) and (3) together. The results of the post-hoc comparisons among the means of the four methods suggest a general tendency toward improvement of the factor estimation by using other personality variables. Some notable examples are factors E, H, and O in situation one, factors E, F, M, and O in situation two, factors A, E, G, N, and O in situation three, and factors A and O in situa-

tion four. For the role variables the improvement is not quite so striking, although factor O is improved in situation one and two, factor Q₃ in situation two, and factor H in situation three.

The inclusion of both personality and motivation variables leads to a negligible and insignificant increase in estimation above that gained by personality alone. Two explanations are possible. The first is that the number of predictors is already so high that the information any new variable adds is redundant. A second explanation is that we should expect no great increase in predictability of a personality factor by the inclusion of motivation variables since the two domains have already been shown to be relatively independent (Cattell, et al., 1964). In fact, this seems to be the case, since an inspection of the factor structure matrix showed low zero-order correlations between the personality factors and motivation variables. Although the two domains are apparently factorially independent, it could, of course, be true, as Cattell, et al. (1964) have suggested, that the two combine additively in the prediction of most everyday life behavior. One would certainly expect from all common sense considerations that test-taking could be among the behaviors so affected by motivation. The most likely explanation is surely that we have not yet succeeded in designing the objective motivation batteries to center upon the motivation operative in the role situation.

REFERENCES

- Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- Cattell, R. B. *Personality and motivation structure and measurement*. New York: World Book, 1957.
- Cattell, R. B. Personality, role, mood, and situation-perception: a unifying theory of modulators. *Psychological Review*, 1963, 70, 1-18.
- Cattell, R. B. The Scree test for the number of factors. *Multivariate Behavioral Research*, 1966, 1, 245-276.
- Cattell, R. B. Trait view theory of perturbations in ratings and self-ratings (L(BR)- and Q-data): its application to obtaining pure trait score estimates in questionnaires. *Psychological Review*, 1968, 75, 96-113.
- Cattell, R. B. and Digman, J. M. A theory of the structure of perturbations in observer ratings and questionnaire data in personality research. *Behavioral Science*, 1964, 9, 341-358.
- Cattell, R. B., Eber, H. W., and Tatsuoka, M. *Handbook for the*

- Sixteen Personality Factors Questionnaire*. Champaign, Ill.: Institute for Personality and Ability Testing, 1969.
- Cattell R. B., and Horn, J. L., Radcliffe, J. A., and Sweney, A. B. *Handbook for the Motivation Analysis Test*. Champaign, Ill.: Institute for Personality and Ability Testing, 1964.
- Cattell, R. B. and Muerle, J. L. The 'maxplane' program for factor rotation to oblique simple structure. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1960, 20, 569-590.
- Cattell, R. B., Radcliffe, J. A., and Sweney, A. B. The nature and measurement of components of motivation. *Genetic Psychology Monographs*, 1963, 68, 49-211.
- Cornbach, L. J. Response sets and test validity. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1946, 6, 475-494.
- Eber, H. W. and Cattell, R. B. Maximizing personality scale validities on the 16 PF by the computer synthesis service. Champaign, Ill: IPAT, 1966.
- Edwards, A. L. *The social desirability variable in personality assessment and research*. New York: Dryden, 1957.
- Harris, C. W. Formula for the significance of a factor loading. Manuscript for private circulation, Educational Psychology Department, University of Wisconsin, 1965.
- Hays, W. L. *Statistics for psychologists*. New York: Holt, 1963.
- Hundleby, J. D., Pawlik, K., and Cattell, R. B. *Personality factors in objective test devices*. San Diego, Knapp, 1965.
- Hurley, J. R. and Cattell, R. B. The Procurstes program: producing direct rotation to test a hypothesized factor structure. *Behavioral Science*, 1962, 7, 258-262.
- Wiggins, J. S. Personality structure. *Annual Review of Psychology*. 1968.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

DIFFERENCES BETWEEN THE MILLER ANALOGIES TEST SCORES OF PEOPLE TESTED TWICE

JEROME E. DOPPELT
The Psychological Corporation

THE Miller Analogies Test (MAT) is a measure of scholastic aptitude that is widely used by universities as one of the bases for selecting graduate students and by business and government as an aid in hiring high-level personnel. The test is administered in licensed Centers which exercise strict control over the materials. The names of all examinees and their scores are reported by the Centers to The Psychological Corporation, the publisher of the MAT. At the examinee's request, his score (or scores) on the MAT will be sent to specified institutions or individuals. Each year a number of people take the MAT for the second or third time. This study is concerned with the scores of people who were tested twice.

Although those who are retested with the MAT constitute a small percentage of the total number tested, they now number more than 1,000 persons per year. Individuals take the MAT more than once for various reasons. There are those who feel that their first score is not truly indicative of their abilities and hope to improve their score on the second testing. Some who were tested several years ago may prefer to supply a current score with an application to a school or for employment. In some situations, admissions or personnel officers have suggested to applicants that they take the MAT again. Such suggestions are frequently made in order to obtain confirmation of either low or high scores. And finally, some people may simply have forgotten they were tested with the MAT in the past and unknowingly apply for a retest.

From the point of view of the official who must evaluate an applicant's MAT score, several questions regarding the results of retesting would seem relevant. How stable are the MAT scores? What kinds of differences are found when scores on first and second testings are compared? How does the time interval between testings relate to the difference between scores? Does the administration of the same form twice yield results which differ from those obtained when different forms are given in the two testings? What is the relationship between initial scores and differences between first and second scores?

Answers to many of these questions have been sought by earlier investigators. The results of two studies published about ten years ago are of interest. Spielberger (1959) analyzed the data from three small samples (total $N = 48$) of psychology students and reported significant average gains between first and second testings, of approximately five points. Different forms of the MAT were used in the two testings. The time intervals were not specified for all cases, but they appear to have been relatively short. Spielberger also studied improvement according to score on the initial testing. He divided his cases into three initial-score groups and reported that the "effects of practice on the MAT scores were most facilitative for Ss with low initial scores and the amount of improvement was inversely related to initial MAT scores." The range of initial scores among Spielberger's cases was 45 to 88 with a mean above 70, indicating a relatively superior group.

Coladarci (1960) studied the MAT scores of 56 male candidates for administrative positions in education. These men had taken the MAT twice with a median time interval of 15.8 months between testings. The product-moment coefficient of correlation between initial and retest scores was .82. A difference of 7.4 points between the means of the two testings was found. This was attributed to the experience intervening between the two administrations rather than to any practice effect, "in view of the modal intervals of time involved." No significant correlation was found between score gains and the time between testings. Coladarci also studied the relationship between initial score and improvement score after classifying his subjects into three groups according to initial-score level. He reported that "in our sample there was

no relationship between the magnitude of the initial score and the amount of improvement, either with obtained scores or estimated true scores." Coladareci noted that this finding was in disagreement with that reported by Spielberger and suggested that this may be a result of the different types of samples in the two studies. The range of initial MAT scores in the Coladareci study was 15 to 82 with a mean of 43.

The present study sought answers to the questions raised above by analyzing the data of two large twice-tested samples selected from the publisher's registry of scores. The first sample, which includes 1,690 cases collected during 1959 and 1960, will be called the 1960 Sample; the second sample, consisting of 624 cases, most of whom were tested for the second time in 1968, will be identified as the 1968 Sample. It cannot be assumed that the samples studied in this paper are representative of all persons who are tested with the MAT, but it is reasonable to suppose that they are representative of persons who voluntarily or by request apply for a retest.

The basic data consisted of the score from each administration of the MAT, including the form¹ of the test given, the difference between the second and first scores, and time in months between testings. Each sample was divided into those who were tested with the same form of the MAT on both occasions and those who were tested with different forms. For these groups, correlations were computed among the scores, time intervals, and score differences. The coefficients, along with the corresponding means and standard deviations, are shown in Table 1 for the 1960 and the 1968 samples.

In recent years an effort has been made to avoid administering as a retest the form of the MAT that had been given as the first test. To this end, one form of the MAT was set aside to be used only for retesting. The result of this procedure may be seen in the smaller per cent of persons retested with the same form in the 1968 Sample (20%) than in the 1960 Sample (41%).

In the 1960 Sample, the "same form" and "different forms"

¹ Forms G, H, J, and K were administered to the subjects of the 1960 Sample; Forms H, J, K, L, and R were administered to the 1968 Sample. To make Form G scores comparable to scores on other forms, two points were subtracted from the Form G scores in the range 32-72, as recommended in the MAT Manual.

TABLE 1

Intercorrelations among MAT Scores from First and Second Testings, Difference between Scores, and Time between Testings

	First MAT Score	Second MAT Score	Score Diff. (Second Minus First)	Time Interval (Months)	Mean	SD	N
1960 Sample							
First MAT Score		.87	-.21	.17	44.3	16.4	997
Second MAT Score	.89		.30	.10	50.3	16.8	
Score Difference	-.15	.31		-.12	6.0	8.5	
Time Interval	.03	.03	.00		15.6	17.8	
Mean	45.3	51.9	6.6	25.7			
SD	15.9	16.5	7.5	19.1			
N 693							
1968 Sample							
First MAT Score		.86	-.25	.09	43.7	14.4	501
Second MAT Score	.82		.29	.10	51.9	14.6	
Score Difference	-.14	.45		.01	8.2	7.8	
Time Interval	.03	-.01	-.06		8.6	8.7	
Mean	34.8	43.7	8.9	12.0			
SD	13.9	15.5	9.0	9.2			
N 123							

Note.—For each sample, data above the diagonal are based on cases tested with different forms of the MAT on the two occasions; data below the diagonal are based on cases tested with the same form of the MAT on both occasions.

groups have very similar MAT means and standard deviations for each testing. The second test score is higher, on the average, by about six to seven points. In the 1968 Sample, the first testing mean score of the "same form" group is much lower than that for the "different forms" group. There seemed to be no apparent reason for this finding. It is seen as a sampling variation, associated with the relatively small size of the former group in comparison with the latter group. The average gains in score, however, are similar for the two groups, and about two points greater than the differences found in the 1960 Sample.

The average time between testings in the 1960 Sample is longer for the "same form" group (25.7 months) than for the "different

forms" group (15.6 months). Table 1 shows that the time between testings in the 1968 Sample is considerably shorter than it was in the earlier sample. Although the time interval for the "same form" group is still found to be longer than it is for the "different forms" group, both figures are about half of what they were for the corresponding groups in 1960. Thus, retesting with the MAT took place sooner, in the 1968 Sample, than it did about eight years earlier.

The correlation between first and second testings may be regarded as a coefficient of stability for the MAT. For the 1960 Sample, coefficients of .89 and .87 are shown in Table 1 for the "same form" and "different forms" groups, respectively. For the 1968 Sample the coefficients for the corresponding groups are .82 and .86. The coefficient of .82 was obtained for the smallest of the four groups, the 123 people of the "same form" group in the 1968 Sample. The standard deviation of first testing scores for this group was about 13 per cent smaller than the standard deviation of the "same form" group in the 1960 Sample (13.9 as compared with 15.9). This type of restriction may account, in part, for the lower coefficient that was obtained in the later sample. In general, the stability coefficients are high and similar to the alternate form reliability coefficients reported in the MAT Manual. This finding is notable when it is recognized that the average time between testings ranges from 8.6 months to 25.7 months, over the four groups.

The correlations between the scores on either the first or second testing and the time interval between testings do not indicate strong relationships. For the four groups in the two samples the coefficients range from $-.01$ to $.17$. Since people apply for a retest with the MAT for quite different reasons, this finding might have been anticipated.

The correlations of the *difference* between scores (second score minus first score) and the time interval between testings are also reported in Table 1 for each group in the two samples. Here, as is true of the correlations of either first or second scores with time interval, the coefficients are low and are unimportant from a practical viewpoint. Further study of the relationship between score difference and time is summarized in Table 2.

The time between testings was divided into six intervals (shown

TABLE 2

Means and Standard Deviations of MAT Score Differences According to Time Interval between Testings*

Form		Time Interval in Months						Total Group
		0-2	3-6	7-12	13-24	25-36	37+	
1960 Sample								
Same	<i>N</i>	49	75	116	139	111	203	693
	Mean	8.6	5.9	6.7	6.1	6.3	6.7	6.6
	<i>SD</i>	8.2	7.3	8.0	7.2	7.1	7.3	7.5
Different	<i>N</i>	301	126	169	168	99	134	997
	Mean	7.1	6.1	6.0	6.0	6.0	3.8	6.0
	<i>SD</i>	8.7	7.6	8.3	7.3	9.9	9.0	8.5
1968 Sample								
Same	<i>N</i>	17	21	37	35	11	2	123
	Mean	10.9	7.7	9.4	8.2	9.8	—	8.9
	<i>SD</i>	5.8	8.7	10.7	8.7	8.1	—	9.0
Different	<i>N</i>	190	73	83	145	6	4	501
	Mean	8.4	8.3	7.3	8.2	—	—	8.2
	<i>SD</i>	7.9	7.3	8.2	7.5	—	—	7.8

Note.—Mean and SD were not computed when *N* was less than 10.

* Second testing score minus first testing score.

in Table 2), and the mean and standard deviation of score differences within each interval were computed. The average difference between scores for the total "same form" group is only slightly greater than that between the "different forms" group in each sample. There is some indication in both samples that those retested with the same form show higher gains when the time between testings is less than three months. Although the differences between the "same form" and "different forms" groups are generally small, it nevertheless seems desirable to give a different form of the MAT when retesting. People who are retested shortly after their first testing probably have an advantage if the identical form is administered in the second session. Furthermore, the

use of a different form, regardless of the time between testings, is more consistent with the idea of an independent evaluation than a repeat testing with the same form.

In Table 1 the highest coefficients, aside from the coefficients of stability, are found between score differences and second testing scores, followed by the correlations between score differences and first testing scores. There is not much value in detailed study of score changes in relation to the score on the *second* testing since the chronology of the situation is not relevant to practical usage. It may be helpful, however, to consider the relationship between the score on the first testing and the change in score after retesting.

As a practical guide to the interpreter of MAT retest data, a single table which indicates expected changes in score on retest, according to the initial test score, was prepared. Table 3 provides such information based on the combined data from the 1960 and 1968 samples. The table shows the 75th, 50th, and 25th percentiles of the distribution of score differences, for various score levels on the first testing. With the exception of those who scored 70 or higher on the first testing, at least 75 per cent of the group at each

TABLE 3

Selected Percentile Equivalents of the Distribution of Differences between First and Second MAT Testings According to Score on First Testing 1960 and 1968 Samples Combined*

First Testing Score	N	Range of Differences	Score Difference Percentile			Per Cent of Differences Which Were		
			75th	50th	25th	Positive	Negative	Zero
80-100	36	-26 to 13	4	0	-3	50	39	11
70-79	112	-10 to 19	9	4	0	73	23	4
60-69	263	-28 to 22	9	4	1	75	21	4
50-59	419	-16 to 27	11	6	1	77	18	5
40-49	508	-23 to 33	12	7	2	79	16	5
30-39	470	-16 to 32	13	8	2	82	14	4
20-29	415	-13 to 44	13	7	2	81	15	4
Below 20	91	- 8 to 34	15	9	4	86	10	4
Total	2,314	-28 to 44	12	7	1	79	17	4

Note.—The table is read as follows: Scores between 80 and 100 on the first testing were obtained by 36 people; the range of the differences between scores on the second and first testings was from -26 to 13. In this distribution of differences the 75th percentile was 4 points, i.e. about 25 per cent of the group obtained second testing scores that were more than 4 points higher than their first testing score. The 50th percentile of the distribution of differences for this group was 0 or no change; the 25th percentile was -3, indicating that the bottom quarter of the distribution of differences for this group showed a loss of 3 or more points on retesting. Fifty per cent of the differences were positive, reflecting a gain on the second testing; 39 per cent of the differences in this group were negative, indicating lower scores on retesting; and 11 per cent obtained the same scores on retesting (zero difference).

* Second testing score minus first testing score. A negative difference means a lower score on the second testing and a zero difference indicates no score change.

score level increased their scores on the second testing. (Examinees who scored 70 or higher on the MAT have demonstrated very superior test performance. One would expect fewer increases in score among people who achieved such high scores on the first testing.) Nevertheless, it should be kept in mind that a substantial number of people obtained scores on the second testing which were lower than their first testing scores. Table 3 shows that the "per cent of differences which were negative" varies from 10 to 39 when the examinees are classified by score on first testing.

Although the range of differences between second and first MAT scores is enormous (-28 to 44), the middle 50 per cent of differences is contained within a band of 11 points (12 to 1). The average difference between test and retest scores is a gain of about seven raw score points. Evidence of a regression effect may be seen. Examinees whose initial MAT scores are between 50 and 80 gain about five points, on the average, when tested a second time. Examinees with initial test scores below 50 show an average gain of approximately eight points, on retesting. The people in the very high-scoring group (those with scores of 80 or higher) show an average gain close to zero, while those in the lowest category of first testing score (below 20) show an average gain of nine points. In general, a gain in excess of 12 points is likely to be found in about one fourth of the group who are tested twice.

Discussion

The writer must admit to a certain amount of fascination with test-retest data such as those provided in this report. It is usually difficult to obtain the scores of a large number of people who have been retested after a considerable period of time. Various types of classifications can be made for study of the scores and score differences, and some of these data have been presented in the foregoing tables. But then one is faced with the nagging, practical question: Now that we have these results, to what end can the admissions officer or the employer use them? It is at this point that the writer's fascination with the data changes to concern, because no simple, clear answer may be given. Some suggestions are offered here.

If both scores of an individual are below the range imposed by the institution's policy, or if both scores are above it, there is no

serious problem in deciding which score to accept. However, if one score falls below a critical level and the second is above that level, a careful review of the situation is indicated.

As a practical matter, it may be assumed that score differences after a short time, say less than one year, are more associated with measurement error than with important changes in the examinee. Large differences over a short time interval may reflect an invalid testing on either the first or second occasion due to illness of the examinee, poor test administration or the like. When in doubt, a third testing might be requested. When the scores straddle a critical range, the wisest course would be to consider other information about the individual rather than routinely to accept either the first or second score or to average them.

Over relatively long periods of time, intervening experience such as education may have contributed to a large gain in score. (It is possible, too, that a loss in score could be related to events in the time between testings.) When a long time has elapsed between two testings, the second score is more likely than the first to reflect accurately the current status of the examinee.

It is obvious, but nevertheless important, that the score on a test should not be the sole basis for an important decision, such as admission to graduate school or employment by an organization. The individual's previous record of accomplishment and any other relevant information must be seriously considered. The results of retesting should be evaluated judiciously. Determining how the test scores complete the picture offered by other available evidence is the most appropriate method for arriving at a satisfactory decision.

Summary

A study was made of the scores on the Miller Analogies Test of people who were tested on two different occasions. Two large samples of cases were selected from the publisher's files. These cases may be considered representative of those who requested a retest with the MAT at the time the data were collected. In the first sample, the retesting was completed no later than 1960, while in the second sample, the retesting of most cases was done in 1968.

The basic findings from the two samples are very similar. The stability coefficients for the MAT range from .82 to .89, with an

average gain of about seven points between testings. The correlations of the time interval between testings with the first or second scores, or with the difference between scores, are low and unimportant from a practical viewpoint. Only small differences were found between the average gains of those tested with the same form twice and the gains of those who were tested with different forms. As a practical matter, retesting with a different form is recommended. The study of gains showed some evidence of a regression effect, with the larger average gain obtained by those with lower initial scores.

A table which shows selected percentile equivalents of the distribution of score differences, according to level of initial test score, was prepared from the combined data of the two samples. This table should provide general information to those interested in the effects of retesting with the Miller Analogies Test. A discussion of the problem of interpreting score differences has been included.

REFERENCES

- Coladarci, A. P. An analysis of Miller Analogies Test score changes. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1960, 20, 817-823.
- Spielberger, C. D. Evidence of a practice effect on the Miller Analogies Test. *Journal of Applied Psychology*, 1959, 43, 259-263.

ELECTRONIC COMPUTER PROGRAM AND ACCOUNTING MACHINE PROCEDURES

WILLIAM B. MICHAEL, Editor
University of Southern California

JOAN J. MICHAEL, Assistant Editor
California State College, Long Beach

<i>Setwise Regression Analysis—A Stepwise Procedure for Sets of Variables.</i> JOHN D. WILLIAMS AND ALFRED C. LINDEM	747
<i>A Note on Generating Multivariate Data with Desired Means, Variances, and Covariances.</i> J. R. CAPRA AND R. S. ELSTER	749
<i>A Ten Factor Unequal "N" Analysis of Variance Program.</i> NORMAN K. RUBIN AND ALAN L. GROSS	753
<i>A Computer Program for Nonparametric Post Hoc Multiple Comparisons.</i> JAMES J. ROBERGE	755
<i>A Program of Scheffé's Method.</i> RANDALL M. PARKER	761
<i>A Computer Program for the Compilation of Data from Classroom Observation Systems Having Mutually Exclusive Categories.</i> THOMAS B. GREGORY	763
<i>An Item Analysis and Scoring Program for Summated Rating Scales.</i> RICHARD L. KOHR	769
<i>The Use of the Common/Data Statement to Determine the Type of an Event in Simulation Studies.</i> EDWIN L. ANDERSON	771
<i>An EDP System Package for Scoring the Interpersonal Check List.</i> DONALD E. LANGE	775
<i>MERMAC Test and Questionnaire Analysis System.</i> LAWRENCE M. ALEAMONI	777

SETWISE REGRESSION ANALYSIS—A STEPWISE PROCEDURE FOR SETS OF VARIABLES

JOHN D. WILLIAMS AND ALFRED C. LINDEM

The University of North Dakota

SETWISE regression analysis is a new technique developed by the authors to allow a stepwise solution when the interest is in sets of variables rather than in single variables. Thus, the setwise regression procedure bears a strong resemblance to the stepwise regression procedure. There are, however, advantages to be gained by the use of setwise regression analysis, and a disadvantage of the stepwise procedure is overcome.

A disadvantage of the usual stepwise procedure is that it becomes inappropriate when there are more than two categories being binary coded. A simple example can be made with religious affiliation. Four categories might be used: Catholic, Protestant, Jewish, and Other. Three binary predictors can be made with the first three religious affiliations, and the fourth category can be represented as not having membership in the first three categories. If religious affiliation were used in conjunction with other information, the stepwise procedure would not yield a valid indication of the importance of the religious variables. The setwise procedure, on the other hand, would allow a direct approach to such a situation.

The setwise procedure drops one *set* of variables at a time in a stepwise fashion. There will be as many steps as there are sets. The steps are accomplished by an iterative procedure that allows the R^2 (multiple correlation coefficient squared) term to be maximized at each step in a backward stepwise procedure. Once a set is discarded, the set is no longer considered at later steps. One set is discarded at each step, until there is only one set remaining.

Input

Data cards contain for each observation the criterion and predictor variables in any format or order. Parameter cards specify problem identification, number of observations, total number of variables, number of sets, criterion variable, optional printout of data, and optional printout of residuals. Set selection cards specify the number of predictors in a set, and the variables included in a set.

Limitations

The maximum dimensions are as follows:

99,999 observations, 40 variables including the criterion variable, and 10 sets of predictor variables.

Computer and Program Language

The program is written in FORTRAN IV level F for the IBM 360 (64K).

Output

The sets of variables remaining in the prediction equation are given for each stage. Also given for each stage is an analysis of variance for the regression, the beta and regression weights, the means and standard deviation for each variable, R^2 , R , $1 - R^2$, and the loss in the R^2 term for each stage.

A printout of the program and sample output will be supplied on request.

A NOTE ON GENERATING MULTIVARIATE DATA WITH DESIRED MEANS, VARIANCES, AND COVARIANCES

J. R. CAPRA AND R. S. ELSTER

Naval Postgraduate School
Monterey, California

THE problem to be discussed involves creating a set of n observations on p variables, with the p variables having specified means, variances, and covariances. The method which will be presented differs from those previously given by Kaiser and Dickman (1962) and Wherry, Naylor, Wherry, and Fallis (1965), in that the procedure does not use the models of principal component or factor analysis.

Derivation and Procedure

Let A be a p by n matrix of n independent observations on p variables. If the p variables are independent, with means of zero and unit variances, then, assuming the normal model A is distributed as a sample from a multivariate normal population with a mean vector of 0 and a variance-covariance matrix equal to the identity matrix. More succinctly, A is distributed as $N(0, I)$.

Anderson (1958, p. 21) showed that if one transforms A in the following way:

$$Z = CA, \quad (1)$$

then Z is distributed as $N(0, CIC')$ or $N(0, CC')$, where C is a p by p matrix used to transform A . Given a specified correlation matrix R , the problem is to decompose R such that $CC' = R$, where C will be a lower triangular matrix since R is symmetric.

If one can derive C and if one has specified the correlation matrix R , then a transformation exists which, when applied to A , will

give a set of observations on p variables with means of zero, unit variances, and the specified correlations among them. It is then a simple task to apply linear transformations in order to achieve the desired means and variances.

The numerical technique for deriving C from R uses Crout factorization (Kunz, 1957, pp. 226-229). The following recursion, which is easily programmed, allows C to be derived (Odell and Feiveson, 1966):

$$\begin{aligned} C_{ii} &= R_{ii}/\sqrt{R_{ii}}, & 1 \leq i \leq p \\ C_{ii} &= \sqrt{R_{ii} - \sum_{k=1}^{i-1} C_{ik}^2}, & 1 < i \leq p \\ C_{ij} &= \left[R_{ij} - \sum_{k=1}^{i-1} C_{ik}C_{jk} \right] / C_{ii}, & 1 < j < i \leq p \\ C_{ij} &= 0, & i < j \leq p. \end{aligned} \quad (2)$$

Ordinarily, a researcher will establish the n observations on each of the p variables of A by using a random number generator. The assumption made is that the random number generator will yield variables having zero means, unit variances, and zero intercorrelations. Because random number generators yield these characteristics only in the limit, a researcher may wish initially to adjust A such that in fact the p variables do have zero means, unit variances, and zero intercorrelations. (Of course, this operation will somewhat distort the randomness of the sample.) Nevertheless, to allow this adjustment, if the user judges it to be desirable, the following option is available.

Let X be a $p \times n$ matrix obtained from the random number generator, the sample mean vector being given by \bar{x} and the sample variance-covariance matrix being given by M . The goal is to transform X into a set of data with sample means of zero and a sample variance-covariance matrix of I . By definition,

$$XX' = M.$$

Consequently, what is needed is a matrix, D , such that

$$DXX'D' = DMD' = I.$$

This matrix could then be used to transform the matrix X into one having a sample variance-covariance matrix I :

$$Y = DX.$$

Now, since

$$DMD' = I,$$

then

$$M = D^{-1}D'^{-1},$$

which means we can obtain D^{-1} by Crout factorization and D itself by simple matrix inversion.

The sample means of this new set of observations are given by

$$\bar{y} = D\bar{x},$$

where \bar{y} is a p component vector. If Y_i refers to the i th column of Y and \bar{y}_i refers to the i th element of the \bar{y} vector, a matrix with zero means can be obtained by subtracting \bar{y}_i from each element in Y_i for each i , from 1 to p .

Required Input Data

All that is required of the user are the correlation matrix, the desired mean and variance for each variable, and the sample size (number of observations on each variable) which he desires to have generated. Of course, the user should insure that the correlation matrix is nonsingular and positive semidefinite.

Output from the Program

The program will generate a multivariate sample from a population with the specified means, variances, and covariances. Sample means, variances, and correlation coefficients are also computed.

Summary

A method is shown for creating a set of n observations on p variables, with the p variables having specified means, variances, and covariances. This method differs from previous techniques in that it uses Crout factorization to develop the desired variance-covariance matrix instead of using the methods of component or factor analysis. Because the procedure assumes that it begins with p variables having zero means, unit variances, and zero intercorrelations, a procedure is also given for transforming the original data so that they fulfill these conditions.

REFERENCES

- Anderson, T. W. *An introduction to multivariate statistical analysis*. New York: Wiley, 1958.
- Kaiser, H. F. and Dickman, K. Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika*, 1962, 27, 178-182.
- Kunz, K. S. *Numerical analysis*. McGraw-Hill, 1957.
- Odell, P. L. and Feiveson, A. H. A numerical procedure to generate a sample covariance matrix. *American Statistical Association Journal*, 1966, 61, 199-203.
- Wherry, R. J., Sr., Naylor, J. C., Wherry, R. J., Jr., and Fallis, R. F. Generating multiple samples of multivariate data with arbitrary population parameters. *Psychometrika*, 1965, 30, 303-313.

A TEN FACTOR UNEQUAL "N" ANALYSIS OF VARIANCE PROGRAM

NORMAN K. RUBIN AND ALAN L. GROSS

The City University of New York

THIS program will perform an analysis of variance for a wide class of experimental designs: complete factorial, nested, randomized blocks, split plots, and other similar designs.

Unlike many of the commonly available Analysis of Variance programs, this procedure possesses the following characteristics:

1. The only restriction on cell size is that no cell is vacant. Cells may contain differing numbers of observations. There is also no upper bound to the number of observations.

2. The program in its present form will process designs having up to $k = 10$ factors.

3. The only restriction on the number of levels allowed for the i th factor (L_i) is that the product $\prod_{i=1}^k (L_i + 1)$ be less than the dimension of the array DATA. For example, to process a $5 \times 9 \times 3 \times 5 \times 2$ design the dimension of Data must be at least $6 \times 10 \times 4 \times 6 \times 3 = 4320$. On an IBM 1130 with 16K of memory; the dimension of data is normally 5000.

4. The program, which is written in ASA BASIC FORTRAN, uses no external files. Thus there should be no problem in adopting the program to a computer having a FORTRAN compiler. A section of comments in the program listing describes in detail the changes to be made.

5. The program allows as an option the user to insert a FORTRAN subroutine to preprocess or transform the input data.

The unequal "N" design is analyzed using an approximate method described by Scheffé (1959). Basically this method leads one

to do an analysis of variance of the cell means with an adjustment to the final sum of squares. When the cell sizes are equal, the program produces the exact solution.

Program input consists of a single control card, variable format card(s), and the data. The data for each cell are preceded by a header card giving a cell ID number and the number of cell observations. The cells themselves can be entered in any order.

Output consists of the normal ANOVA table properly labeled, cell means, cell standard deviations, and the cell sizes.

A source copy or listing of the program may be obtained from either author.

REFERENCE

Scheffé, H. *The analysis of variance*. New York: Wiley, 1959.

A COMPUTER PROGRAM FOR NONPARAMETRIC POST HOC MULTIPLE COMPARISONS

JAMES J. ROBERGE¹

Temple University

Most researchers in the behavioral sciences follow the decision to reject a null hypothesis, on the basis of an F -test, with a post hoc analysis of specific linear contrasts, using one of the various multiple comparison procedures (e.g., Scheffé, 1953, 1959; Tukey, 1953). Recently, Nemenyi (1963), Dunn (1964), and Rosenthal and Ferguson (1965) discussed similar procedures which may be employed following the rejection of the null hypothesis by a given nonparametric test, i.e., the Kruskal-Wallis (1952) one-way analysis of variance for rank data or the Friedman (1937) two-way analysis of variance for rank data. The program described in this paper is designed to perform these nonparametric post hoc multiple comparisons.

Rationale

Nemenyi (1963) proposed a procedure for determining which pairwise comparisons among k treatment populations are significant. This procedure requires the calculation of the statistic d , and the value of the constant C , for each pair of samples.

The value of d is calculated by the following formula:

$$d = |\bar{R}_i - \bar{R}_{j'}|$$

where \bar{R} is the mean rank for a given sample (or experimental condition) and j and j' represent indices of disjoint subsets of the k samples (or experimental conditions).

¹The author gratefully acknowledges the support for this research which was provided by a Faculty Research grant funded by Temple University.

The constant, C , for a given comparison is calculated by one of the following formulas:

If the Kruskal-Wallis one-way analysis of variance is the appropriate statistical technique, then

$$C = \sqrt{\chi^2_{\alpha, k-1} \cdot \left[\frac{N(N+1)}{12} - \frac{\sum_{s=1}^r (t_s^3 - t_s)}{12(N-1)} \right] \cdot \left[\frac{1}{n_j} + \frac{1}{n_{j'}} \right]}$$

where k is the number of independent samples, χ^2 has $k - 1$ degrees of freedom, $\alpha = .05$, N is the total number of observations (or ranks), r is the number of sets of tied observations, t is the number of tied observations for a given set s , n is the number of subjects in a given sample, and j and j' are as defined above; on the other hand, if the Friedman two-way analysis of variance is the appropriate statistical technique, then

$$C = \sqrt{\chi^2_{\alpha, k-1} \cdot \left[\frac{k(k+1)}{6n} - \frac{\sum_{s=1}^r (t_s^3 - t_s)}{6n^2(k-1)} \right]}$$

where k is the number of experimental conditions, n is the number of subjects (or matched subjects), and χ^2 , α , r , s , and t are as defined above.

According to Nemenyi's test, the hypothesis that two samples j and j' were drawn from identically distributed populations is rejected if the value of d for the samples exceeds the value of C .

Dunn (1964) presented a procedure whereby rank sums from a combined ranking of k independent samples (Kruskal-Wallis model) are used to determine which populations differ. This procedure requires the calculation of the statistic y/σ for each comparison. Each of these values is then compared with tabled values for the standard normal distribution to determine approximate probability levels.

The components of the statistic y/σ are calculated by the following formulas:

$$y = \frac{\sum_i T_i}{\sum_i n_i} - \frac{\sum_{i'} T_{i'}}{\sum_{i'} n_{i'}}$$

where y is an arbitrary contrast, T is the rank sum for a given sample, n is the number of subjects in a given sample, and j and j' are as defined above;

$$\sigma = \sqrt{\left[\frac{N(N+1)}{12} - \frac{\sum_{i=1}^r (t_i^2 - t_i)}{12(N-1)} \right] \left[\frac{1}{\sum_j n_j} + \frac{1}{\sum_{j'} n_{j'}} \right]}$$

where σ is the standard deviation of an arbitrary contrast and N , r , s , t , n , j , and j' are as defined above.

Rosenthal and Ferguson (1965) described a procedure which can be employed to construct post hoc confidence intervals for experiments involving n rankings of k objects (Friedman model). This procedure requires the calculation of the mean and standard error of the mean for *each* contrast, and a constant for *all* contrasts.

The mean, \bar{T} , and standard error of the mean, s_T , for each contrast are calculated by the following formulas:

$$\bar{T} = \sum_{i=1}^n \frac{T_i}{n} \quad s_T = \sqrt{\frac{n \sum_{i=1}^n T_i^2 - \left(\sum_{i=1}^n T_i \right)^2}{n^2(n-1)}}$$

where T is the weighted sum of the ranks for a given subject i (or group of matched subjects) and n is the number of subjects (or matched subjects).

The constant, C , for all contrasts is calculated as follows:

$$C = \frac{(k-1)(n-1)}{(n-k+1)} F_{\alpha; k-1, n-k+1}$$

where k is the number of experimental conditions, n is the number of subjects (or matched subjects), F is the usual statistic, and $\alpha = .05$.

The confidence intervals for the various contrasts are of the following form:

$$\bar{T} - \sqrt{C} s_T \leq L \leq \bar{T} + \sqrt{C} s_T$$

where L is an arbitrary contrast and \bar{T} , C , and s_T are as defined above.

Input

The job deck set-up for each analysis is as follows:

Problem card

- Columns
- 1 = Nonparametric test (1 = Kruskal-Wallis; 2 = Friedman)
 - 2 = Nemenyi's test (1 = yes; 0 = no)
 - 3 = Dunn's test (1 = yes; 0 = no)
 - 4 = Rosenthal and Ferguson test (1 = yes; 0 = no)
 - 5-6 = Number of samples or experimental conditions (k)
 - 7 = All possible pairwise comparisons (1 = yes; 0 = no)
 - 8-10 = If column 7 is 0, then the number of comparisons is punched in these columns; otherwise, they are left blank.
 - 11-16 = If column 1 is 2, and column 4 is 1, then the F -ratio for the Rosenthal and Ferguson test (see above) is punched in these columns (Note: the decimal must be punched); otherwise, they are left blank.

Contrasts matrix format card

This F -type variable format card describes each row of the arbitrary contrasts matrix. This format may be punched in any of the columns on the card. If column 7 on the problem card is 1, then this card is omitted.

Arbitrary contrasts matrix

This matrix is entered one row at a time. Each row must begin on a new card and must have k weights indicating the contrast to be made. These cards must be punched in accordance with the F -type contrasts matrix format card (see above). If column 7 on the problem card is 1, then these cards are omitted.

Data format card

This F -type variable format card indicates the location of the raw scores (or ranks) on the data cards. This format may be punched in any of the columns on the card.

Sample Card(s)

A card (or cards) indicating the size(s) of the sample(s). For the Kruskal-Wallis test, the number of subjects in each sample is punched

on the card(s) using 2613 format. For the Friedman test, the number of subjects in the sample (or matched samples) is punched on the card using I3 format.

Data deck

These cards contain the data for each sample (or experimental condition) and must be punched in accordance with the format specified on the *F*-type data format card (see above). For the Kruskal-Wallis test, the data are punched by sample with the data for each sample beginning on a new card. For the Friedman test, the data are punched by subject (or group of matched subjects) with the data for each subject (or group of matched subjects) beginning on a new card.

Last card

If the user wishes to terminate the program, then the card immediately following the data deck must have the word FINISH punched in columns 1 to 6. However, if the user wishes to analyze another set of data, then this card is a blank card and the job deck is arranged sequentially (as described above) beginning with the problem card.

Output

The computer output for the Kruskal-Wallis test includes the value of H (corrected for tied ranks), the number of degrees of freedom, and the average rank for each sample. Moreover, if the null hypothesis is rejected, then the output for Nemenyi's test includes the values of d and C for each contrast, and the output for Dunn's test consists of the value of each contrast, the standard deviation of each contrast, and the value of the statistic y/σ .

The computer output for the Friedman test includes the value of χ^2 (corrected for tied ranks), the number of degrees of freedom, and the average rank for each experimental condition. Furthermore, if the null hypothesis is rejected, then the output for Nemenyi's test is as described above, and the output for the Rosenthal and Ferguson test consists of the 95 percent confidence interval for each contrast.

Capabilities and Limitations

The program, which is written in FORTRAN IV, can handle a maximum of 30 samples (or experimental conditions) and 200

subjects per sample (or experimental condition). Jobs may be run sequentially as described above.

Availability

Copies of this paper and a source listing which includes input and output data for sample problems can be obtained by writing to Dr. James J. Roberge, Temple University, Department of Educational Psychology, Philadelphia, Pennsylvania 19122.

REFERENCES

- Dunn, O. J. Multiple comparisons using rank sums. *Technometrics*, 1964, 6, 241-252.
- Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 1937, 32, 675-701.
- Kruskal, W. H. and Wallis, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 1952, 47, 583-621.
- Nemenyi, P. B. *Distribution-free multiple comparisons* (Doctoral dissertation, Princeton University). Ann Arbor: University Microfilms, 1963. No. 64-6278.
- Rosenthal, I. and Ferguson, T. S. An asymptotically distribution-free multiple comparison method with application to the problem of n rankings of m objects. *British Journal of Mathematical and Statistical Psychology*, 1965, 18, 243-254.
- Scheffé, H. A method for judging all possible contrasts in the analysis of variance. *Biometrika*, 1953, 40, 87-104.
- Scheffé, H. *The analysis of variance*. New York: Wiley, 1959.
- Tukey, J. W. The problem of multiple comparisons. Unpublished manuscript, Princeton University, 1953.

A PROGRAM OF SCHEFFÉ'S METHOD

RANDALL M. PARKER

The University of Texas at Austin

SCHEFFÉ (1959) has reported a method for making any or all possible comparisons among means or combinations of means in equal or unequal n ANOVA designs, regardless of whether significant F values are obtained. Because of its flexibility, Scheffé's method is a particularly desirable multiple comparison procedure.

SCHEFFÉ is a FORTRAN program that computes Scheffé's method for up to 100 means or combinations of means. A unique feature of this program is that it computes a corrected F value and the exact probability of F , which is analogous to the usual procedure of computing confidence intervals, but somewhat easier to interpret. The input parameters are the number of group means to be compared, the number of subjects, the error term (usually MS within), the tabled F value at appropriate degrees of freedom and desired probability level (optional), the group means, the n s for each group, and the coefficients for each comparison. Output includes the arithmetic difference among the means, the smallest statistically significant difference among means (only if the F value parameter is input), a corrected F value, and the exact probability of the F value for each comparison. Subroutines PRTS and PRBF (Veldman, 1967) are called by SCHEFFÉ.

REFERENCES

- Scheffé, H. A. *The analysis of variance*. New York: Wiley, 1959.
Veldman, D. J. *Fortran programing for the behavioral sciences*. New York: Holt, Rinehart, and Winston, 1967.

A COMPUTER PROGRAM FOR THE COMPILATION OF DATA FROM CLASSROOM OBSERVATION SYSTEMS HAVING MUTUALLY EXCLUSIVE CATEGORIES¹

THOMAS B. GREGORY²
Indiana University

THE past decade has witnessed the development of a profusion of observation instruments that attempt to record objectively the ongoing verbal and/or nonverbal events occurring in the classroom. Flander's Interaction Analysis (1960) and the several revisions of the OScAR (e.g., Medley and Mitzel, 1958; Medley, Schluck, and Ames, 1968) are notable examples of this trend. Simon and Boyer (1967) reported that over 50 such systems had been developed by 1967. Even modest extrapolations of the accelerating pace at which such systems are continuing to be produced leads one to conclude that well over 100 probably exist today.

The rapid growth in popularity of classroom observation instruments may be partially explained by identifying their two basic functions. First, the instruments can provide a pre-service or in-service teacher with valuable descriptive feedback about his teaching which can, in turn, lead him to attempt alternative tactics perceived as being more desirable. Second, such observation systems can be valuable criterion measures in research applications.

Many computer programs exist which compile the data required to fulfill either or both of these functions for individual observa-

¹ This research was conducted with partial support of USOE Grant No. OE 6-10-108, Research and Development Center for Teacher Education, The University of Texas at Austin.

² The author wishes to acknowledge the assistance of Dr. Donald J. Veldman in the preparation of a preliminary form of this program.

tion systems. However, two problems arise from such an approach. First, many systems are in an almost constant state of evolution. A computer program written specifically for such a system must be rewritten as its system changes. Second, many systems are devised for a specific circumstance which may never occur again. Developing programs for such unique situations becomes prohibitive because of the time factor involved.

Program *Cosan* (Classroom Observation System ANalysis) is a general purpose FORTRAN program that can compile data needed for fulfilling either or both the feedback and/or research functions for any observation system containing up to 25 mutually exclusive categories. A series of subroutines allows a wide range of input and output options that facilitate adaptation of the program to diverse feedback and/or research contexts.

Data Deck Arrangement

Card 1

Parameter Control Card

Columns:

- 1-2 Number of categories in the system (Max., 25).
- 3-4 Number of ratios to be specified (Max., 20). A 0 suppresses this function.
- 5-6 Number of behavior sequences to be punched (Max., 20). A 0 suppresses this function.
- 7-8 Percentage desired for minimum cut-off point on high-frequency cell listings (e.g., a 3 will cause only those cells containing at least 3% of the total behavior to be listed). A 0 suppresses this listing.
- 9-10 A 1 if *any* printed output is desired; 0 otherwise.
- 11-12 A 1 if *any* punched output is desired; 0 otherwise.

The following 4 parameters may be left blank if cols. 9-10 = 0.

- 13-14 A 1 if printed matrix is desired, 0 to suppress it.
- 15-16 A 1 if printed column (category) totals are desired, 0 to suppress them.
- 17-18 A 1 if printed percentages of total behavior for each category are desired, 0 to suppress them.
- 19-20 A 1 if printed ratios are desired, 0 to suppress them.

The following 3 parameters may be left blank if cols. 11-12 = 0.

21-22 A 1 if punched column (category) totals are desired, 0 to suppress them.

23-24 A 1 if punched percentages of the total for each category are desired, 0 to suppress them.

25-26 A 1 if punched ratios are desired, 0 to suppress them.

Punched data are returned 10 variables per card. Each card contains the subject's identification field (cols. 1-8) a data identification code (col. 9 for which T = column totals, P = percentages of total behavior for each category, R = ratios and S = sequences), and a data card number (col. 10). For example, a punched card displaying R2 in cols. 9-10 indicates that it is the second ratio card for the subject identified in cols. 1-8.

Graphic Display Spacing Specifications

The following 7 parameters allow the optional graphic display of the sequence of behavior to reflect logical groupings of categories within the system. For example, Flander's original 10 category system contains 4 logical groups: indirect influence (4 categories), direct influence (3 categories), student talk (2 categories), and silence (1 category). Therefore, 4, 3, 2, and 1 would be the first 4 parameters which would be followed by 3 parameters left blank.

27-28 Number of categories desired in first group. Set equal to category *N* if no grouping is desired. Set equal to 0 to suppress the graphic display.

The following six parameters may be left blank if cols. 27-28 equal either the category *N* or 0.

29-30 Number or categories desired in a 2nd group if needed, blank otherwise.

- 31-32 Number or categories desired in a 3rd group if needed, blank otherwise.
- 33-34 Number or categories desired in a 4th group if needed, blank otherwise.
- 35-36 Number or categories desired in a 5th group if needed, blank otherwise.
- 37-38 Number or categories desired in a 6th group if needed, blank otherwise.
- 39-40 Number or categories desired in a 7th group if needed, blank otherwise.

Card 2 Category Code Card

Columns:

- 1-1 FORTRAN character on data cards corresponding to category 1.
- 2-2 FORTRAN character on data cards corresponding to category 2.
- ⋮

$N-N$ FORTRAN character on data cards corresponding to category N .

$N + 1-N + 1$ FORTRAN character on data cards used as a stop signal indicating the end of a subject's data.

The ordering of the output display of categories may be altered by simply arranging the order of their corresponding FORTRAN characters on this card. A / cannot be used as a category symbol.

Next N_r Cards Ratio Specification Card(s) (N_r = Parameter Control Card, cols. 3-4). Omit if $N_r = 0$.

Columns:

- 1-20 Ratio identification. Any legal FORTRAN characters may be used.
- 21-80 Ratio specification. Categories should be identified by the same symbols used on data cards. Addition is assumed between categories except when the beginning of the denominator is signaled through use of a /. One / is permitted, though not mandatory, in each "ratio." A category may be "weighted"

by repeating it a desired number of times in the numerator and/or denominator. No blanks are permitted in this field until the end of the ratio. The total of all categories may be indicated by using the stop signal (see Category Code Card).

Next Card

Sequence Specifications Card

Omit if Parameter Control Card, cols. 5-6 = 0.

Columns:

- 1-2 Matrix row number containing cell of first behavior sequence to be punched. (e.g., 4 for the 4, 10 cell.)
- 3-4 Matrix column number containing same cell (e.g., 10 for the 4, 10 cell).
- 5-80 Specification of up of 19 additional behavior sequences (matrix cells) using same format.

Data are returned as the percentage of total behavior located in each specified cell.

Next Card

Beginning of Subjects' Data

Columns:

- 1-8 Subject identification. Any legal FORTRAN characters may be used. This field must not be blank.
- 9-10 This field is not read by the machine and is for clerical uses such as numbering the data cards for each subject.
- 11-80 First 70 behaviors (category entries) for this subject.

Repeat same format for second through *N*th cards for subject 1. Any number of behaviors (max. = 1000; i.e., 14 cards plus 20 cols. of card 15) are permitted for each subject. When data for subject 1 are completed, the stop signal should appear in the next data column. Begin subject 2's data on next card.

Last Card

Blank card indicating the end of the data.

Listings of Program Cosan may be obtained either from the Research and Development Center for Teacher Education at The

University of Texas at Austin, Austin, Texas 78712; or from the Author, 125 School of Education Building, Indiana University, Bloomington, Indiana 47401.

REFERENCES

- Flanders, N. A. Interaction analysis in the classroom. Minneapolis: University of Minnesota, College of Education, (Pre-Publication draft), 1960. (mimeographed).
- Medley, D. M. and Mitzel, H. E. A technique for measuring classroom behavior. *Journal of Educational Psychology*, 1958, 49, 86-92.
- Medley, D. M., Schluck, C. G., and Ames, N. P. *Assessing the learning environment in the classroom: A manual for users of OSAR 5V*. Princeton, N. J.: Educational Testing Service, 1968. 47 pp. (mimeographed).
- Simon, A. and Boyer, E. G. (Eds.), *Mirrors for behavior*. Philadelphia: Research for Better Schools, Inc., 1967.

AN ITEM ANALYSIS AND SCORING PROGRAM FOR SUMMATED RATING SCALES¹

RICHARD L. KOHR
Bucknell University

THIS program has three major options for processing data originating from an attitude scale of the Likert-type (summated rating scale). Option one produces an item analysis following procedures similar to those outlined by Edwards (1957, pp. 152-154). Option two yields a printout and/or punch cards containing the results of various scoring operations. Option three indicates that the user desires both options.

Input to the program are item scores and any desired coded information (e.g., demographic variables) which the user wishes to have reproduced, along with the scale score(s), on the punch card output.

Regardless of the options selected, certain information is supplied to the user. This includes a page summarizing the options requested and the information contained on the program control cards. A second page gives the following total scale information: total number of items, number of response choices or categories, sample size, mean, variance, standard deviation, unbiased estimate of population variance, estimated standard deviation, Coefficient Alpha, standard error of measurement, and the estimated average inter-item correlation.

The item analysis option provides the following. For *each* item a table is printed for both a *low total* attitude score group (lowest 27%) and a *high total* attitude score group (highest 27%). The tables include the frequency and proportion of occurrence of each

¹The development of this program for the IBM System 360/67 was supported by the Pennsylvania State University Computation Center.

response choice. The *item* mean and standard deviation is presented for both contrast groups. In addition, the output includes an adjusted correlation between each item and the score based on the composite of the remaining items on the scale. This represents the correlation between an item and the total score with the contribution of that item to the total score removed (Guilford, 1953). This correlation is based on *all* the respondents and not just those in the high/low contrast groups. Lastly, a frequency distribution of the total attitude scores is printed along with the *N*, mean, variance, and standard deviation of the two contrast groups.

The scores on which the various statistics are based are the values assigned to the response choices which must form a meaningful continuum. Often this takes the form of strongly agree, agree, . . . , strongly disagree. Since attitude scales frequently contain some statements which are favorable toward the object and some unfavorable, it is necessary to reverse the scoring of certain items. A conventional scoring system assigns the higher value to the response indicating favorableness such that when the item scores are summed they are directionally consistent. The program can perform this reversing operation and thereby can permit considerable flexibility as to the nature of the input data.

The second option may consist of printed and/or punch card output consisting of each respondent's identifying information (e.g., subject and demographic codes) and total (or subscale) score(s). It is also possible to receive punch cards containing scored items (made directionally consistent) as well as all items pertaining to a subscale clustered together.

The program written entirely in FORTRAN IV (H-level) contains no machine specific subroutines. Write-ups, listings of the source program and trial data, and a copy of the output may be obtained by writing to the author.

REFERENCES

- Edwards, A. L. *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, 1957.
- Guilford, J. P. The correlation of an item with a composite of the remaining items in a test. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1953, 13, 87-93.

THE USE OF THE COMMON/DATA STATEMENT TO DETERMINE THE TYPE OF AN EVENT IN SIMULATION STUDIES

EDWIN L. ANDERSON
Oregon State University

THE determination of times between events in simulation studies based on stochastic processes involves the use of a random number generator (RNG), an expected value for the time of the event which is usually derived from a real system, and the inverse transformation of the cumulative distribution function of some known distribution. When this routine indicates that it is time for a decision to be made concerning an event, a simulation program written in FORTRAN can accomplish this (when the probability of such an event occurring is known) by IF statements and the RNG. For example, if 80 per cent of the students entering a high school counseling center desire to have a conference with a counselor, the RNG can be called to generate random variates uniformly distributed between 0 and 1. If the random variate is .80 or less, the program will branch to a sequence which determines the type of conference and the length of the conference. If the random variate is greater than .80, the program will branch to a sequence which determines the purpose for the student being in the system and the length of time spent in the system.

Input

If it has been decided that the next student entering the system desires a conference, IF statements may be used to determine the type of conference. However, when the number of possibilities is great, this routine becomes somewhat lengthy, and the COMMON/DATA statement in the computer program handles the decision

more efficiently. If the conference types and the probabilities from the real system are identified as listed in Table 1, the COMMON/DATA statement would be:

COMMON/DATA/TYPE(100)
DATA ((TYPE(I),I=1,100)=10(1),2(2),13(3),
18(4),10(5),8(6),33(7),3(8),3(9))

The COMMON/DATA statement can be thought of as the cumulative frequency table displayed in Table 2.

The RNG, supplied with an initial random number, will generate a random variate uniformly distributed between 1 and 100. The cumulative table is searched to locate the position of that variate and the corresponding number representing a type of conference is determined. This is the type of conference for one student. The process would then be repeated for each student entering the system and desiring to have a conference. It is not necessary to input additional random numbers, as the next random variate to be used will be generated by the RNG.

Output

In simulation studies dependent on stochastic processes, the ordinary procedure is to make several computer runs and to use the average of the runs for the simulated system. The data presented in Table 3 are a result of three runs using the COMMON/DATA statement given above.

If the mean probabilities presented in Table 3 simulate the type of conferences held during one day at the counseling center, a simu-

TABLE 1
*Types of Conferences and Probability of Occurrence**

Number	Type	Probability
1	Academic	.10
2	Attendance	.02
3	Employment	.13
4	Personal-social	.18
5	Post High School	.10
6	Records	.08
7	Schedule	.33
8	Vocational	.03
9	Other	.03

* Types of conferences established by six counselors at a high school counseling center. Probabilities were derived by tracing 2,127 students through the counseling center during nine random sampling days during the winter months of 1969-70.

TABLE 2

Cumulative Frequency Table from COMMON/DATA Statement

Random variate	Conference type
98-100	9
95-97	8
62-94	7
54-61	6
44-53	5
26-43	4
13-25	3
11-12	2
1-10	1

lation of several days will result in probabilities which will be a "good fit" to the actual probabilities in Table 1. Thus, the COMMON/DATA statement may be a vital aspect of a simulation program.

TABLE 3

Probability of Type of Conference Using a RNG and COMMON/DATA Statement—100 Counselors

Run 1		Run 2	Run 3	Summary statistics for all runs		
(Initial random no.)		(27935)	(6761)	(69831)		
Conference type	Probability	Probability	Probability	Mean	SD	SE
1	.10	.09	.11	.1000	.0100	.0058
2	.01	.03	.03	.0233	.0115	.0067
3	.15	.13	.18	.1533	.0252	.0145
4	.18	.17	.14	.1600	.0265	.0153
5	.08	.12	.10	.1000	.0200	.0115
6	.07	.06	.05	.0600	.0100	.0058
7	.37	.33	.34	.3467	.0208	.0120
8	.01	.03	.04	.0267	.0153	.0088
9	.03	.04	.01	.0267	.0153	.0088

AN EDP SYSTEM PACKAGE FOR SCORING THE INTERPERSONAL CHECK LIST

DONALD E. LANGE

University of Victoria

ALTHOUGH the Interpersonal Check List (ICL) (LaForge and Suczek, 1955) has found a wide and varied usage, many potential users are reluctant to include it in their assessment battery because of the laborious task involved in keying for all twenty variables and the subsequent computation necessary to obtain its summary scores. To correct this situation, an EDP system package has been prepared which will relieve the ICL user from such clerical labor. The package is offered free of charge to anyone wishing a copy, provided that it is not used for commercial gain (e.g., using the package to set up a center to score ICLs for profit).

Description of the System

The system has been designed to relieve the user from the clerical tasks involved in keying and scoring the ICL form IV (LaForge, 1963). The computer program itself consists of a main driver and two subroutines; it is written in basic FORTRAN IV. Currently, tested packages are available for IBM system 360 OS/DOS, IBM 1130, and PDP 10. If the user wishes to employ the package on other systems, it is a simple matter to change job control language in order to make it compatible, provided of course, that the other system will support standard Fortran IV.

Data input, which may be of two types, may include up to eight response protocols from a single subject. First, the preferred type is from the Document No. 511 standard form IBM 1230 Optical Mark Scoring answer sheets, upon which the testee has marked his responses to the ICL. This standard form is then read on the IBM

1230, which transfers the marks, in special 1230 code, to punch cards that will be later keyed and scored by computer. Second, if the user has already hand-keyed the ICL for the necessary twenty variables, the variable data may be entered for computation of the summary scores. This step, of course, does not utilize the package's ability to key the ICL and to save the user from such an uninteresting task, but it will save him from the lengthy job of computing the summary scores. Therefore, if someone is planning to administer the ICL, it is recommended that he change his answer sheets to Document No. 511 standard form IBM 1230 and ask his examinees to mark the first answer choice for each ICL adjective that applies (leaving it blank if it does not). He should let the item numbers on the answer sheet correspond to the ICL's adjective numbers. Then on receipt of the system package, these protocols may be read on the IBM 1230, keyed, and scored by computer.

Since the output may be either printed, or of printed and punched form, the user is allowed to utilize the resulting ICL data for further analyses without the additional key punching step.

Materials

The materials which will be returned to those who request the ICL Scoring Package will consist of a control sheet and drum card for the IBM 1230 Optical Mark Scoring Reader, the computer program source decks, and a detailed manual on how to utilize the scoring system. When ordering, one should state the configuration of the computer system so that the most nearly applicable package may be returned.

Please send request for this package to Donald E. Lange, Department of Psychology, University of Victoria, Victoria, British Columbia, Canada.

REFERENCES

- LaForge, R. Research use of the ICL. *Oregon Research Institute Technical Report*, 1963, 3.
LaForge, R. and Suczek, R. R. The interpersonal dimension of personality: III. An interpersonal check list. *Journal of Personality*, 1955, 24, 94-112.

MERMAC TEST AND QUESTIONNAIRE ANALYSIS SYSTEM

LAWRENCE M. ALEAMONI

University of Illinois

A test and questionnaire analysis system was designed to assist instructors in developing valid and reliable tests and to provide rapid and meaningful feedback to the instructor and students.

Description of the System

MERMAC is made up of two sets of programs: (a) utility (data manipulation) program, and (b) test and questionnaire analysis programs. The seven utility programs allow the user to copy, edit, match, merge, sequence, sort, and recode the input data. Generally, the purpose of these programs is to prepare the data for input to the test and questionnaire analysis programs. The six test and questionnaire analysis programs allow the user to:

1. Score item data and produce up to forty subscores for each individual. Each item and response may be weighted to arrive at the scores. Any item may be included in more than one subscore and be weighted differently in each.
2. Take scores for a group of individuals and produce a frequency distribution and histogram, mean, median, standard deviation, Kuder-Richardson reliability, standard error of measurement, and Spearman-Brown prophecy for a reliability of .90. In addition, individual raw scores, standard scores, and percentiles may be listed. Individual raw scores and standard scores can be weighted, summed, and the sum assigned a letter grade. All these data can be easily provided to the student.
3. Return to each student a page containing his test score and

a list of the items he missed with his responses and the correct responses.

4. Analyze his item data by providing a plot of the percentage of individuals responding to the keyed response by fifths of the total score distribution. For each item alternative the proportion of individuals responding, a point biserial correlation, and the number responding to each alternative by fifths is provided.
5. Analyze his item data by using some external criterion rather than the keyed test score.
6. Summarize item data from questionnaires or tests with no known correct answers by providing a frequency distribution of responses, a weighted mean, and a standard deviation for each item. In addition, subscores may be generated with means, standard deviation, split-half reliabilities, and percentage of individuals responding to the contributing items. It is also possible to assign deciles to the item and subscore means based on a table look-up.

Summary

The MERMAC system is written in Basic Assembly Language (BAL) for both IBM System/ 360 models 40 and above and IBM System/ 370 models 135 and above which have Operating System (OS) with Queued Sequential Access Method (QSAM) support.

Additional information about the program and its availability may be obtained from Lawrence M. Aleamoni, Measurement and Research Division, 307 Engineering Hall, University of Illinois, Urbana, Illinois 61801.

BOOK REVIEWS

MAX D. ENGELHART, Editor
Duke University

HENRY MOUGHAMIAN, Assistant Editor
City Colleges of Chicago

<i>Best's Research in Education.</i>	DENNIS M. ROBERTS	781
<i>Blackman and Goldstein's An Introduction to Data Management in the Behavioral and Social Sciences.</i>	JOHN L. WASIK	783
<i>Ferguson's Statistical Analysis in Psychology and Education.</i>	LEWIS R. AIKEN, JR.	785
<i>Gagne' and Gephart's Learning Research and School Subjects.</i>	JOHN A. R. WILSON	787
<i>Isaac and Michael's Handbook in Research and Evaluation.</i>	DENNIS M. ROBERTS	790
<i>McCall's Fundamental Statistics for Psychology.</i>	LEWIS R. AIKEN, JR.	792
<i>Robinson's Heredity and Achievement.</i>	ROBERT A. GORDON ...	793
<i>Steger's Readings in Statistics for the Behavioral Sciences.</i>	GERALD M. GILLMORE	799

John W. Best. *Research in Education*. (2nd ed.) Englewood Cliffs, N. J.: Prentice-Hall, 1970, Pp. vi + 399. \$8.95.

It has been 11 years since Best's first edition of *Research in Education* and judging from the extent of the current revisions, it doesn't look as though much has happened during this time span. To be fair, however, I should preface further remarks by saying that the book appears to be a good one and will meet the reasonable objectives set by Best. He says that the book won't make one an expert in research—and, *that* is refreshing to hear. As Traub (1969) has discussed, facilitating the process of making better consumers of research as compared to becoming a researcher are two quite different tasks calling, perhaps, for different learning approaches. But this strays from the immediate mission. What follows is a brief run-down of the contents of this second edition with what I hope are relevant comments interjected along the way.

Best covers 13 chapters and four appendices in approximately 400 pages. The first three chapters deal with such questions as what research is, identification of a research problem (which includes about five pages of suggested topical areas for research), and the use of reference materials. The chapter on reference materials—while possibly useful—takes up about 50 pages by listing far too many sources of information that students could go to. In addition, Best goes over the Dewey decimal system and other points about proper use of the library. While this material may provide some leads that graduate students can follow up—(a) the necessity for the extensiveness of the list of reference material is highly debatable and (b) the inclusion of "how to use the library" information seems inappropriate for a text designed primarily to cater to graduate students. True—some graduate students might still not know how to use the library—however, these people must be beyond hope.

Chapters 4, 5, and 6 respectively deal with historical, descriptive, and experimental research. In each case, the author points out the major purposes of each of these research approaches along with examples and appropriate limitations. The most notable difference between the revised and earlier edition is the reworking of the chapter on experimental methods. Best brings in, summarizes and discusses quite effectively notions from Campbell and Stanley (1966).

Matching studies are down played this-go-round as compared to the '59 version. To me, this chapter is the nicest aspect of his current reworking of the book.

Chapter 7 discusses various tools of research such as questionnaires, surveys, psychological tests, interviews, etc. However, included is a fairly inadequate presentation of the concepts of reliability and validity. These topics could have definitely been improved and elaborated in more detail. It would not have hurt (although it might have been necessary to have placed it after the chapters on data analysis) to identify some of the correlational ways reliability and validity can be estimated and to discuss some of the factors that affect them.

Chapter 8 is very brief and discusses ideas related to the interpretation of data. It includes such things as scales of measurement (with an inappropriate example of ranks of professors for the nominal scale), ways of tabulating data and—interesting enough—a filing scheme by McBee Systems for coding and retrieving information.

Chapters 9 and 10 present basic computational techniques—nine dealing with descriptive data analysis and 10 following on to inferential ideas. This is a reorganization, expansion and splitting up of material covered in one chapter in the earlier edition. The current treatment is logically better but still has some bugs that could have been ironed out. For example, the assumed mean procedure is outlined with the statement made that the assumed mean value is the *true* mean value. This is misleading since that would only occur in a perfectly symmetrical distribution. There are also formulas for computing the median, percentiles and the p *th* percentile from grouped data. All these seem rather cumbersome and unnecessary—especially for a book of this type. In addition, the interquartile and semi-interquartile ranges are discussed along with the formula for finding the standard deviation using the assumed mean procedure. Again—these will confuse the student by putting unnecessary clogs in the wheel. The mechanics of how to compute the standard deviation are given without much indication of “why” and “how” it is used. As for correlation, the r and ρ formulas are presented without much mention of why one would logically use one as opposed to the other. Regression lines are discussed in one section while the prediction equation itself is given in another section. Bringing these two together would have been better. In the chapter on inferential statistics, Best discusses various sampling techniques along with such things as the standard error of the mean, critical ratio, t test, simple ANOVA and chi square. In general, this material is presented well and in a sensible format.

Chapter 11 deals with the research report and its write-up. Again, while the material is sensible and follows traditional patterns of

stylistic considerations, most of the material is covered locally at a particular university where they publish their own style guides. However, for those who are simply interested in basic mechanics and hints of the format of a research report (especially those who aren't in university settings), the chapter will probably be helpful.

Chapter 12 presents several summaries of "significant" research studies along with limited discussion of various points of these studies. While I'm dubious of the usefulness of this chapter, my feeling is only a hunch.

The last chapter deals with various aspects of federal support for educational research. Regional labs, the ERIC system, the support (although it is now dwindling!) for the training of educational researchers, and ways and means for preparing moneyseeking research proposals are discussed. This is one of the best chapters in the book.

Finally, in terms of the content, the four appendices include (1) a research report evaluation form, (2) a glossary of statistical formulas and symbols, (3) areas under the normal curve, and (4) percentile scores for any given two scores.

In summary, Best has produced an improved edition of *Research for Education*. The basic improvements include the chapters in experimental research, data analysis reorganization, and information of federal support to educational research. The book is quite readable with a low difficulty level and should be useful in assisting students toward their pursuit of an overview to aspects of the research process. More thought should have been given to the descriptive analysis chapter, but, other than this, Best's book is "better" than it previously was and should be well received.

REFERENCES

- Campbell, D. T. and Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally & Co., 1966.
- Traub, R. E. Review of Gephart and Ingle's *educational research: selected readings*. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1969, 29, 1010-1013.

DENNIS M. ROBERTS
Department of Psychology
East Carolina University

Sheldon Blackman and Kenneth M. Goldstein. *An Introduction to Data Management in the Behavioral and Social Sciences*. New York: John Wiley, 1971, pp. vi + 104. \$7.95 and \$5.95 (Paperback).

This brief treatment of data setup and analysis procedures for use with computers is intended to be used by persons with no previous exposure to computers and/or electronic data processing equip-

ment. A major theme of the text is that a plan for the conduct of an experiment requires a parallel plan for the use of a computer in data handling and statistical analyses. Jargon has been kept to a minimum and a glossary of words commonly encountered in reading about data management and program writeups has also been provided.

After an introductory chapter, nine other chapters are presented with content related to translation of data from a raw to a machine readable form and use of "canned" computer statistical program packages.

Within the data translation section of the text, there are short discussions relating to principles for coding data and the presentation of data in a matrix arrangement (Chapter 2); the development of a data-processing format and the punching of data into cards (Chapter 3); the concept of a variable format (Chapter 4); the editing and coding of data prior to card punching (Chapter 5); and of the purposes and use of auxiliary data processing equipment (Chapter 6).

The discussion of canned statistical programs and/or systems begins in Chapter 7 with a short description of the purpose and examples of computer installation control cards (*i.e.*, JCL's—Job Control Language cards). Chapters 8 and 9 present a description of and the commonly required program cards required in setting up BMD and P-STAT program runs. This presentation contrasts the use of the variable format (*i.e.*, BMD) and free-format (*i.e.*, P-STAT) methods of specifying data to be processed in a computer run. Chapter 10 covers a wide range of topics related to acquisition and modification of other available statistical programs and packages. Two appendices are also included. Appendix A presents a specimen set of data in a coded matrix arrangement while Appendix B discusses matrix manipulations useful for data reduction purposes. Both could have been omitted with no loss in communication of content to the readers. Appendix A presents information previously given in the body of the text while the second appendix does not present enough information for an individual without a background in measurement to be able to use the ideas presented within this section. Questions are provided at the end of each chapter, however, they are lacking in structure and are unlikely to help a person who is in a self-study situation to understand the basic concepts presented in the chapter. In contrast, a short list of readings, keyed by topic, appears at the end of each chapter and is likely to be more helpful.

The cost of the paperback version text indicates that a purchaser pays dearly especially when one considers that the 100 pages contain little in the way of mathematical equations, generally the cause for higher prices on scientific books. However, a casual perusal of

several popular textbooks on behavioral and social science research methods indicates students will find little help from these sources on how to set up data and program cards in order to make a computer run. Moreover, it is this reviewer's experience that students learn to use a computer for data management and analysis purposes if he has had a course in computer programming (e.g., FORTRAN) or by having access to a person with experience in using a computer. If the novice computer user does not find himself/herself in either of these two situations, then this text would seem likely to offer a way of getting his/her data through a computer run without being overwhelmed by the process of having to use a computer. In conclusion, while this text is not likely to be purchased by persons with substantial experience in conducting research, it would seem appropriate for use by persons who are setting up data to run through a computer for the first time.

JOHN L. WASIK

Center for Occupational Education
North Carolina State University

George A. Ferguson. *Statistical Analysis in Psychology and Education*. (3rd ed.) New York: McGraw-Hill, 1971. Pp. xii + 492. \$10.95.

There appears to be a general tendency for textbooks that go through several editions to increase in size with each new edition. Unfortunately, this accretion is seldom the result of clearer explanations and more examples, but rather an attempt to incorporate more topics within the book. Ferguson's statistics book is illustrative of this tendency, and also the inflationary tendency of subsequent editions to increase in price. The length of the book has increased from 347 pages in the first editions, through 446 pages in the second edition, and now 492 pages in the third edition. The price of the book has gone from \$7.00 in 1959 to \$7.95 in 1966 to \$10.95 in 1971!

Considering the fact that 12 years, and apparently many adoptions, have intervened between the first and third editions, one has a right to expect the author to have done some housekeeping by making corrections and responding to previously noted shortcomings of the book (Binder, 1960; Glass and Maguire, 1966). However, many of the criticisms made by the reviewers of the first and second editions of this book apply equally well to the third edition. Although the present reviewer does not agree with Glass and Maguire (1966) that Ferguson's "prose tends to be dry," a detailed check of the third edition to determine whether the specific problems noted in the earlier reviews had been attended to revealed that many of them had not. Surely Ferguson and his publisher were aware of

these reviews, but perhaps the former would explain, as he did in the Preface (3rd ed.) in regard to the criticisms of the four consulting editors, that "... limitations of time have prevented me from incorporating some of their more insightful recommendations."

Nevertheless, in spite of its datedness in spots (Fisher is not now living, and 1944 is not *currently*!) and its seeming imperviousness to criticism, the third edition remains like the first and second ones a fairly technically correct and at least average introductory statistics book. The book is divided into four parts: I. Basic Statistics (13 chapters), II. The Design of Experiments (7 chapters), III. Nonparametric Statistics (2 chapters), and IV. Psychological Test(s) and Multivariate Statistics (5 chapters).

The first 13 chapters have changed little since the first edition, reflecting in their emphasis on scales of measurement (nominal, ordinal, interval, ratio) the language of psychological statistics during the late 1950s. The eight or so exercises at the end of each chapter are still rather unimaginative in content, but they may serve the purpose. In Chapter 4, s^2 is defined both as $[\sum (X - \bar{X})^2 / N]$ and $[\sum (X - \bar{X})^2 / N - 1]$, but for most of the book the second definition is used. This ambiguity is partly the result of the author's seeming preference for the greater simplicity of $[\sum (X - \bar{X})^2 / N]$ when it is employed in other formulas. In general, the first 13 chapters are straightforward, well-written, and perhaps sufficient for a traditional, one-semester course. To be sure, there are typos and other problems. For example, formula 10.5 (p. 140) needs an \bar{X} in the denominator; in a symmetrical distribution the mean, median, and mode do not necessarily coincide (p. 52), and Ferguson's explanation of sampling is still inadequate. Concerning the material on correlation, McNemar (1969) and Glass and Stanley (1970) handle the topic better, although the latter book is replete with typos, misprints, and other minor errors. Ferguson has also scattered the correlation material throughout the book (Chapters 7, 8, 21, 23, 24, and 26), whereas most statistics books put more of it all together.

Comparing Part II of the third edition with that of the second edition, Chapter 14 has been moved to 21, Chapter 16 to 25, and Chapters 17-20 down to 14-16. Chapter 17 on three-way ANOVA is new, and Chapter 14 of the second edition (Rank Correlation Methods) has been moved to Chapter 21 in the third edition. Chapter 18 of the third edition (Multiple Comparisons) is a short, new addition which finally mentions Tukey's method. However, the poor statement of assumptions underlying ANOVA and the omission of the independence assumption—which also characterized the two earlier editions—have not been corrected. And in Chapter 16 (p. 243), the repeated measures ANOVA is still not in customary form, the reasons behind using s_{int}^2 in the denominator of the F test being

unclear. Also, is Ferguson serious about the necessity of testing for differences among subjects by means of an F ratio? Finally, the differences between random, fixed, and mixed effects ANOVA models are not explained—a serious omission.

The title of Part IV (Psychological Test and Multivariate Statistics) is somewhat misleading, since discriminant function analysis, MANOVA, and other important multivariate techniques are not discussed. Furthermore, Chapter 25 (Score Transformations: Norms) belongs in Part I, and Chapter 26 (Partial and Multiple Correlation) belongs closer to Chapters 7 and 8. Finally, the reviewer seriously doubts whether Chapters 23, 24, and 25 on test statistics and factor analysis should have been included in this introductory statistics book. The author would have invested his time more wisely by attending to the criticisms of the earlier versions of the book and the suggestions of the consulting editors rather than writing new material which may serve only to pad the book. But as Binder (1960) and Glass and Maguire (1966) concluded in their reviews of the first and second editions, it is an adequate text—no worse than most and better than many.

REFERENCES

- Binder, A. M. Review of George A. Ferguson's *Statistical analysis in psychology and education*. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1960, 20, 863-869.
- Glass, G. V and Maguire, T. O. Review of George A. Ferguson's *Statistical analysis in psychology and education*. (2nd ed.) EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1966, 26, 1075-1078.
- Glass, G. V and Stanley, J. C. *Statistical methods in education and psychology*. Englewood Cliffs, N. J.: Prentice-Hall, 1970.
- McNemar, Q. *Psychological statistics*. (4th ed.) New York: Wiley, 1969.

LEWIS R. AIKEN, JR.
Guilford College

- Robert M. Gagné and William J. Gephart (Eds.). *Learning Research and School Subjects*. Itasca, Ill.: F. E. Peacock. 1968. Pp. ix + 268. \$6.50.

Learning Research and School Subjects is the record of the Eighth Annual Phi Delta Kappa Symposium on Educational Research held October 28 and 29, 1967 at the University of California, Berkeley.

The record of a symposium where leaders in a field meet and present overviews of their work on a topic has become one of the

most useful summaries a student of the topic can obtain. The discussants, who are themselves experts on the topic, bring out facets of the presentation that often remain obscure in other forms of research reporting.

Five chapters corresponding to five sessions make up the book. These are (1) Concept Learning and Concept Teaching presented by Robert Glaser and discussed by Patrick Suppes, Evan R. Keislar, and James J. Gibson. (2) Perceptual Learning in Educational Situations presented by Eleanor J. Gibson and discussed by J. M. Stevens and Henry C. Ellis. (3) Two Scientific Approaches to the Management of Instruction presented by Ernst Z. Rothkopf and discussed by Arthur R. Jensen and Stanford C. Erickson. (4) Three Conceptual Approaches to Research on Transfer of Training presented by M. C. Wittrock and discussed by Lawrence M. Stolurow and Henry C. Ellis. (5) The Quest for Prescriptive Values in Our Educational Programs presented by George C. Thompson and discussed by Winfred F. Hill and Paul H. Mussen.

The focus of the different sections is on moving laboratory research to practice in the classroom. The problems of making this shift are most clearly delineated in the first two chapters, areas in which more definitive research is available than in the later chapters. One of the factors making translation difficult is the need in the laboratory to define goals in measurable terms that can be statistically treated with some assurance. When related phenomena are seen in the classroom, they usually are much more complex, contaminated by indeterminate variables and measurable with little assurance that the changes are related to the names given the variables.

The problems of evaluation are brought out clearly in the first chapter on Concept Learning and Concept Teaching. Suppes pointed out in his discussion that the laboratory studies deal almost exclusively with concept identification in which the learner, animal or human, has to identify whether the critical factor is roundness, blueness, or darkness. These concepts are already within the repertoire of the learner but it is necessary for him to sort out which of them are critical in the pattern changes that are being rung. The number of trials required, the time it takes, or the number of errors made before reaching criterion are all easily measured. The concept is so clear cut that once it is discerned, no reasonable doubt exists that the learner has, in fact, identified this concept and not some other. However, in school, teachers spend most of their time trying to teach children to *form* concepts of democracy, addition, gravity, or phoneme-grapheme correspondence. These are not concepts already within the repertoire of the learner and determining the accuracy and the depth of the formulation is very difficult. The order of the scale on which measurement is made is different from that used

in concept identification. The possibility of error of measurement in the classroom setting is much greater than in the laboratory, but in school learning it is complex concept formulation that is important.

The situation is reminiscent of the story of the drunk down on his knees beneath a lamp post obviously looking for something. A kind passerby, perhaps a psychologist, offered to help and after some questioning learned that a key had been dropped. Further questioning revealed that the key had been dropped half way up the block. To the question of why he was looking here the drunk replied, "The light is better." Hopefully the laboratory psychologists can be enticed away from where the measurements can be clear and precise to where the problems really are.

Eleanor Gibson's presentation of perceptual learning was stimulating. Her idea of perceptual hierarchies from lower order units to higher order units is provocative. She sees perception as a process of filtering the relevant from the noisy rather than a process of adding or associating. In her higher orders she speaks of students perceiving structures and developing strategies of perceptual search. Some question might be raised about the extension of the definition of perception to processes of conceptualization and self-directing activities which result in learning.

Rothkopf spoke of the calculus of practice as a description of the increasingly fine analysis of what is being learned and taught. He speaks of mathemagenic activities as describing the way in which the student reacts in the instructional situation. These activities include translation, segmentation and processing. Rothkopf points out that many of the things learned by students in school "may be quite undesirable from the point of view of the instructor." He describes research designed to bring these activities under more direct and predictable control of the teacher or researcher.

Wittrock describes three approaches to research on transfer of training. These are "(1) simplistic S-R models (2) mediated generalization models (S-r-s-R) and (3) Gagné's model of hierarchical learning sets." These are quite straight forward and grow from one another. One of the participants raised the question as to whether they could not all be subsumed under a single category with subsets. Wittrock agreed that this is possible but expressed his belief that separating the approaches makes further research more likely.

Thompson closed the conference with a presentation on values, their nature, validity, and how they can be taught. Obviously, this is a most complex and difficult topic. The conference reached the same conclusion.

If a criticism should be made of the book and of the conference, it would have to be that the topic was too broad to handle adequately. Even so, a clear understanding of the point of view of

many of the participants does emerge. Students will find *Learning Research and School Subjects* a useful orientation to the views of a number of leaders in a number of fields.

JOHN A. R. WILSON
University of California
Santa Barbara, California

Stephen Isaac and William B. Michael. *Handbook in Research and Evaluation*. San Diego: Robert R. Knapp, 1971. Pp. vi + 186. \$7.95 and \$4.95 (paperback).

Handbook in Research and Evaluation falls somewhere between recipe oriented books like Bruning and Kintz (1968) and Winer (1971) on the one hand, and more discussion-oriented books such as Kerlinger (1964) and Dayton (1970) on the other hand. Isaac and Michael have put together a brief volume that achieves a happy balance between the two ends of the continuum. As they state in the Foreword, the book was prepared for a researcher or research evaluator who simply wants an overview, a summary of alternative approaches, an exhibit of reference models, or a listing of strengths and weaknesses of different methods of research. In doing so, the authors rightfully add the caution that this "balance" approach has risks involved—those being possible oversimplifications of the material. However, according to this reviewer, the material has been handled in such a way as to make the risk factor minimal.

The book is organized into five chapters. Chapter 1 deals with planning research and evaluation studies. Topics range from common mistakes in the formulation of a research problem, through advantages of a pilot study, to planning for computer analysis and data processing. Chapter 2—the lengthiest of the five—deals with research designs and strategies of research. The authors, for convenience sake only, categorize nine different types of research (historical, descriptive, true experimental, etc.). In addition, simple research designs are explicated along with brief discussions of important topics such as confounding, interaction, internal and external validity, statistical regression, and some disadvantages of matching as a control device. Chapter 3 presents information on instrumentation and measurement. A test evaluation form is given, the oft reprinted normal curve table from The Psychological Corporation is presented, techniques for item analysis are discussed, reliability and validity are outlined, along with information on such instrumentation as mailed questionnaires, research interviews, the semantic differential and creativity tests. Chapter 4 provides a summary of the most widely used statistical tools—both basic descriptive and elementary inferential. Computing guides are given for such measures as percentiles, means and standard deviations, correlation, chi square, and the *t* test. This chapter also summarizes informa-

tion concerning hypothesis testing (type I and II errors), the power of a statistical test, and sampling. Chapter 5, the final one, presents criteria and guidelines for writing research proposals (do any of these work?) and reports. Included is a checklist for evaluating a research article, examples of vague and clearer behaviorally written objectives, the inevitable *Taxonomy of Educational Objectives* (Bloom et al., 1956), a model for evaluating school programs, and finally—an excerpt from Skinner (1959) concerning a dissenting view of research methodology and theory. An interesting way to end the book. One wonders if it were a Freudian slip.

A few minor negative points on the material should be pointed out. In Chapter 3 on measurement, it is mentioned that the stanine scale consists of nine intervals—each being one-half of a standard deviation wide. This is true for stanines 2 through 8, but *not* stanines 1 and 9; these go to infinity. A small point, but it could have been clarified. In the section on item analysis, chi square is given as the technique, with the chi square value being computed from the 2×2 table of Right-Wrong versus High scorers-Low scorers. In this discussion, it is pointed out that a significant chi square value indicates that a dependable difference between the number of *correct* answers exists between the high and low scorers, and therefore the item should be retained. However, one can obtain a significant chi square value when more of the low scoring students answer the item correctly than do the high scoring students. Therefore, it should be mentioned that *if* more of the high group get the item correct *and* the chi square value is significant, then one should consider the item to have discriminating power. To the person reading the book, the lack of making this point clear could cause some confusion.

In Chapter 4 on statistics, two points should be mentioned. First, in the computing guide for t , a footnote says "concerning negative numbers, -1.70 is *less* than -1.65 ." (p. 134). However, this is not true as far as the t value is concerned. A t value of -1.70 is *larger* than -1.65 . The minus sign only indicates the direction of the differences in the sample means. In the first computing guide for chi square, the basic formula given for the 2×2 table is the one where each cell is labeled A, B, C or D and then the manipulations are made with sums and multiplications of these values. However, for tables greater than 2×2 , the more traditional expected minus observed frequency formula is then presented. It is my feeling that the more familiar formula should have been presented for all chi square computations. However, the alternate formula could be given as useful when 2×2 tables are being used. I do think though, that the more common formula should be the basic one that is started with, and that giving it will allow the reader to better grasp the basic idea of doing chi square.

With the exception to the few minor points mentioned above, my general reaction to the *Handbook* is very favorable; in fact it seems to me to be the best book covering educational research material to come out for a long time. The major strength lies in the chapter on research design. Isaac and Michael have produced an excellent practical guide for the audience that they intended it for. As someone said, most textbooks could be condensed by at least 50 per cent without any substantial loss in meaning. The current authors have done precisely that with what I consider to be a *gain* in usefulness. I would strongly suggest that people interested in the research endeavor to investigate the contents of this well done book.

REFERENCES

- Bloom, B. S. *Taxonomy of educational objectives, handbook I: cognitive domain*. New York: David McKay Co., Inc., 1956.
 Bruning, James L. and Kintz, B. L. *Computational handbook of statistics*. Glenview: Scott, Foresman and Co., 1968.
 Dayton, C. Mitchell. *Designs of educational experiments*. New York: McGraw-Hill, 1970.
 Kerlinger, Fred N. *Foundations of behavioral research*. New York: Holt, Rinehart and Winston, Inc., 1964.
 Skinner, B. F. *Cumulative record*. New York: Appleton-Century-Crofts, Inc., 1959.
 Winer, B. J. *Statistical principles in experimental design*, (2nd ed.). New York: McGraw-Hill, 1971.

DENNIS M. ROBERTS
East Carolina University

Robert B. McCall. *Fundamental Statistics for Psychology*. New York: Harcourt, Brace Jovanovich, 1970. Pp. viii + 419. \$9.50.

Considering the fact that dozens of elementary statistics textbooks are currently available, it may be viewed as presumptuous to offer yet another one. But Robert McCall's *Fundamental Statistics for Psychology* is a couple of standard deviations above the mean, and it is certainly worth examining by teachers of the introductory course in statistics for the behavioral sciences.

The book has been designed with an eye to teachability and the psychology of learning, an orientation which more textbook writers might well adopt. Although a knowledge of high school algebra and geometry is sufficient mathematical background for 95 per cent of the book, the author recognizes the importance of logic, formulas, proofs, and statistical concepts. A review of symbols, fractions, factorials, exponents, factoring, roots, and interpolation is included in an appendix.

The use of tabular inserts for more technical material and detailed proofs adds to the continuity and ease of reviewing the text. Repe-

tition of symbols and their names, explanations of tables both in the text and at the tables themselves, in addition to isolation and review of concepts and formulas, are other important procedures used by the author to insure understanding and retention.

The 12 chapters of the book include the usual topics in descriptive and inferential statistics at the elementary level: frequency distributions, central tendency, variability, percentiles, regression, correlation, hypothesis testing, and *t* tests. In addition, one-way and two-way analyses of variance are covered in two chapters; nonparametric techniques in a long, 41-page chapter; and further topics in probability in the final chapter. Almost twice as many pages are devoted to the last six chapters, which deal with statistical inference, as to the first six chapters on descriptive statistics. Chapters 7 and 8 on hypothesis testing are particularly well written.

Exercises are placed at the end of a section—another useful pedagogical device—rather than waiting until the end of the chapter. However, the student must turn to an appendix to confirm his answers. Also at the back of the book are the customary statistical tables, a glossary of symbols, and an index. The book is attractively packaged in a blue and white cover. In sum, this newcomer should be a respected competitor on the elementary statistics book market, and one that this reviewer is happy to recommend.

LEWIS R. AIKEN, JR.
Guilford College

Daniel N. Robinson (Ed.) *Heredity and Achievement*. New York: Oxford University Press, 1970. Pp. X + 441. \$4.95 (paperback).

This is a collection of readings intended for an introductory course in behavioral genetics, with emphasis on the especially important issues of intelligence, and of racial differences in intelligence. Since it provides an informative introduction covering concepts in genetics, a course built around this book would not require any previous exposure on the part of students to genetics. The two readings by geneticists Hirsch and Dobzhansky, placed late in the book, are also rich in didactic material, and might well be read first along with the introduction. It would be wise for the psychologist instructor, however, to know a bit more population genetics than the book provides, and for the geneticist instructor to know much more about the race-intelligence controversy, statistics, and psychological measurement.

The initial selections are studies which illustrate the genetics of maze-learning ability, spontaneous activity, avoidance conditioning, and memory in rats or mice. For establishing the basic point of there being a genetic basis for behavior, these papers are

invaluable, if sometimes tiresome with minute experimental detail. These are followed by Gordon Allport, discussing traits, and David Rosenthal on familial concordance by sex for schizophrenia—a long paper so packed with close argument that it will be difficult for most students to follow. This section closes with a paper addressed to its theme, the inheritance of personality, by Gottesman. This is a clean, straightforward piece, although its indexes of heritability may now be somewhat dated. Jensen (1967) has presented a revised formula for heritability and has pointed out that unless corrections for attenuation are used, estimates of heritability are too low. According to Jensen, there also appear to be some peculiarities associated with heritability estimates of personality variables.

In a later section, Beach's call for cross-species comparative research, like an earlier one by Verplanck, focuses attention on profound differences in animal behavior that must be rooted in genetics, and Scott's discussion of critical periods, such as in imprinting, presents many examples of acute genotype-environment interaction of a highly special sort. (Oddly enough, the editor fails to point this out, although he makes frequent mention of such interactions in other contexts.) Unfortunately, the critical periods model needs to be scrutinized carefully before its limited applicability to common differences in intellectual performance is apparent, and neither the article itself—which ends on a seductive note about the possibility of "learning 'not to learn'"—nor the editor provides this.

A major part of the book deals with race and intelligence, both directly and indirectly. The editor distinguishes two camps (p. 3), going back almost thirty years to Boring's terms, the "nativistic" and the "empiricistic." "Environmentalistic" might have been a more neutral term to present to students, who will be unaware of the context of Boring's use of these terms in 1942, and who probably have been socialized to regard "empiricists" as the "good guys." The introduction ridicules lay questions such as, "What fraction of intelligence is determined genetically?" with the help of portentous but cryptic references to gene-environment interaction and gene action (p. 4), instead of training the student to think in terms of heritability of IQ by particular populations in stated environments. Genotype-environment interaction really boils down to a statistical question in calculating heritability, and since available evidence (Jinks and Fulker, 1970; Jensen, 1970a) indicates that this component of the variance in IQ is negligible, it is not the bugaboo that the frequent allusions to it would have us believe. Most likely, these allusions are predicated on interactions that are dramatically apparent but which occur only far outside of the usual range of interest of the environmental variables concerned, for example, when they are lethal to the organism. It is also discouraging to students

to be told, as the editor does, that the nature-nurture problem is a "pseudoquestion," without leading them to think in terms of heritability and components of phenotypic variance. If it were a pseudoquestion, this book would hardly be necessary.

"It would be highly unlikely that very significant differences would exist among races in regard to those characteristics which are vital to survival—for example, 'intelligence,'" the editor states (p. 13), and he quotes geneticists Fuller and Thompson in support, who said, "... it is likely that natural selection tends to oppose the establishment of major heritable behavior differences between races." But who decides what size difference is "very significant" or "major?" One standard deviation in IQ may be trivial on the scale of nature, although of considerable consequence on the scale of human affairs. In these same passages (p. 13), the editor advances some extremely dubious assertions in an attempt to account for phenotypic racial differences—almost as though they were nongenetic—brought about by selection pressure, and ends by suggesting that "the translocation of these racial genotypes to cultures calling for very different forms of intellectual expression could place the racial minority at something of a disadvantage. However... the relocated race would contain genotypes whose norm of reaction surely *allowed* adaptation to the new requirements, even if it *preferred* some slightly different form of expression." This appears to be simply yet another a priori attempt to define any genetic differences that might be established as unimportant, instead of talking about their actual possible magnitude. The use made of "norm of reaction" in this connection strikes me as wishful, as does the vague reference to the adaptable genotypes, without consideration of their relative frequency.

It seems to be the prevailing impression that any geneticist is automatically better qualified—as though some kind of Guardian of DNA—than any social scientist to discuss these issues, although some cultural anthropologists claim that *they* are the ultimate authorities (Diamond, 1962). Accordingly, the editor attempts to trump Jensen by playing the geneticist Hirsch, who has "provided a one-paragraph qualification of the facts and views" of Jensen (p. 222). In this paragraph, Hirsch instructs Jensen, with the help of numerous exclamation points and sarcastic asides, not only about genetics, but also about defining race, heritability, and intelligence, and about the education of the disadvantaged as well. Those personally acquainted with Jensen's careful and thoughtful consideration of all of these issues will recognize the injustice being done here, not just to the scientist, but to science itself.

Robinson follows this with a serious misstatement of fact (p. 223). He says that 25 per cent of the black population exceeds the mean IQ of the white population, whereas the correct value has been given by Shuey (1966, pp. 501-502) as 11 per cent. He trifles

with the problem posed by some northern blacks scoring higher in IQ than some southern whites by ignoring the possibility of selective migration and archly asking whether "genotype changes with latitude?" Treating a supposed association between school expenditures and child's IQ in the same manner, and ignoring the association between SES and IQ as a potential source of spuriousness, as well as the failure of the Coleman Report to find important relations between school variables and pupil achievement, he asks, "Does genotype vary with educational expenditures?" The answer to both questions, of course, is quite possibly, yes. In my opinion, the purpose of an introductory text should be to discuss such issues, not to pose polemical questions left unanswered. A bit further on (p. 223), Robinson reports that monozygotic twins, "reared in very different environments, will reveal *average* IQ differences of fourteen points." However, he omits to state that the very same test showed an average difference of nine points for monozygotic twins reared *together* (Gottesman, 1968), and so only about five points of the fourteen could be attributed to the difference in environments between families. Since a comprehensive review of all studies of IQ differences between identical twins reared apart shows that the grand average difference is only 6.6 points (Jensen, 1970a), the large difference of nine points reported by Gottesman even for identical twins reared *together* suggests, as we might expect, that the IQ test in question (Raven's Mill Hill Vocabulary Scale) was less reliable than the Stanford-Binet or Wechsler-Bellevue, which have been used in other such studies. When Jensen (1970a) pooled the IQ's from the Mill Hill with those from another short IQ test given at the same time, thereby enhancing the reliability of the final IQ, the average difference for these twins reared apart became 6.72, which is quite comparable to values observed in the other major studies of such twins, using longer tests (Jensen, 1970a). In evaluating the average difference in IQ between monozygotic twins reared apart, furthermore, it is always necessary to take into account the component due to measurement error, as reflected in average differences between two testings of the *same individual* with alternate forms; these differences average 4.68 for the Stanford-Binet (Jensen, 1970a). One should also give attention to evidence, reviewed by Jensen (1969a, 1970a), that IQ differences between monozygotic twins seem to be associated with prenatal and other biological influences, rather than with the *social* environment. When all of these considerations are taken into account, Robinson's use of the fourteen point difference is seen to be exceedingly misleading. Yet, this is exactly the kind of "fact" that will stick in students' minds.

Unfortunately, Jensen is not represented in this book. There are, however, excellent readings by Burt, and by Erlenmeyer-Kimling

and Jarvik, on heritability of IQ, which in combination with geneticist Hirsch's statement in his article that separate breeding populations are "almost certain to differ" in relative frequencies of different alleles in their gene pools, could set thoughtful students thinking despite the editor's distractions. The 1960 review of psychological studies of race differences, by Dreger and Miller, is also included. Like their later work, it bends over backwards not to draw any conclusions, and suffers consequently from a nomological shallowness. A selection by Wesman on the definition of intelligence defines it as "the summation of the learning experiences of the individual," thereby receiving the editor's endorsement. Nothing is said about intelligence as the *capacity* to learn or, in Jensen's work (1969a), as abstract reasoning ability. Wesman's definition seems to suggest that we can teach individuals all to be very intelligent, although IQ test performance has proven remarkably resistant to coaching, and the school performance of low IQ children has been equally hard to boost on a permanent basis.

Many readers will be irritated by the number of times that intelligence is placed in quotation marks, or referred to as "*it*," in various places. This adds nothing but mystification. Most will also find Hirsch's attack on the mean, and concern with other parameters such as skewness and variance, to be equally excessive, even for the purpose of discrediting "typological" thinking. The mean, after all, is the statistic that best summarizes all of the observations in the distributions in question, and one-way ANOVA is known to be quite robust for slight differences in variance.

A teacher of behavioral genetics will be able to use this book if he supplements it with other readings so as to balance the picture and remain current. I say this with ambivalence, because it would mean giving wider circulation to Hirsch's article, in which he put words in the mouth of the psychologist Garrett that are sufficiently removed from what Garrett actually said to constitute an act that is at least mildly vicious. Suggested supplementary reading would include "must" papers by Jensen (1967, 1969a, 1969b, 1969c, 1970a, 1970b, 1971), and papers by De Lemos (1969) and Garron (1970), the latter two dealing for a change with nonverbal and quantitative abilities. Students should also be exposed to the work of Lesser, Fifer and Clark (1965), which shows cognitive profiles unique to different ethnic groups, but constant across SES. Palmer (1970) and Lane, and Albee and Doll (1970) have shown some environmental differences that do not make a difference in IQ, including having a schizophrenic parent. Some policy considerations are treated well in Jensen (1970c, 1970d), and Bereiter (1970), and moral issues are sensibly discussed in Bressler (1968), Brues (1964), Ingle (1970), and Scriven (1970). Important topics related to the validity of ability tests for disadvantaged groups are

covered in Stanley (1971) and Sattler (1970). A collection of reading from a different perspective appears in Kuttner (1967), and if one wishes a really sweeping overview by an outstanding geneticist, there is Darlington's (1969) book. Finally, for those who would like to give students a whiff of diatribe, there is Alfert (1969a, 1969b), followed by Jensen's replies (1969d, 1969e).

REFERENCES

- Alfert, E. Comment on: The promotion of prejudice. *Journal of Social Issues*, 1969, 25, 206-211. (a)
- Alfert, E. Response to Jensen's rejoinder. *Journal of Social Issues*, 1969, 25, 217-219. (b)
- Bereiter, C. Genetics and educability: Educational implications of the Jensen debate. In J. Hellmuth (Ed.), *Disadvantaged Child; Volume 3*. New York: Brunner/Mazel, 1970, 279-299.
- Bressler, M. Sociology, biology, and ideology. In D. C. Glass (Ed.), *Genetics*. New York: Rockefeller University Press and Russell Sage Foundation, 1968, 178-210.
- Brues, A. M. Statement on statements on racism. *Current Anthropology*, 1964, 5, 107-108.
- Darlington, C. D. *The Evolution of Man and Society*. New York: Simon and Schuster, 1969.
- De Lemos, M. M. The development of conservation in aboriginal children. *International Journal of Psychology*, 1969, 4, 255-269.
- Diamond, S. Letter. *Science*, 1962, 135, 961-964.
- Garron, D. C. Sex-linked, recessive inheritance of spatial and numerical abilities, and Turner's syndrome. *Psychological Review*, 1970, 77, 147-152.
- Gottesman, I. I. Biogenetics of race and class. In M. Deutsch, I. Katz, and A. R. Jensen (Eds.), *Social Class, Race, and Psychological Development*. New York: Holt, Rinehart and Winston, 1968, 11-51.
- Ingle, D. J. Possible genetic bases of social problems: A reply to Ashley Montagu. *Midway*, 1970, 10, 105-121.
- Jensen, A. R. Estimation of the limits of heritability of traits by comparison of monozygotic and dizygotic twins. *Proceedings of the National Academy of Sciences*, 1967, 58, 149-156.
- Jensen, A. R. How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 1969, 39, 1-123. (a)
- Jensen, A. R. Reducing the heredity-environment uncertainty: A reply. *Harvard Educational Review*, 1969, 39, 449-483. (b)
- Jensen, A. R. Intelligence, learning ability and socioeconomic status. *Journal of Special Education*, 1969, 3, 23-35. (c)
- Jensen, A. R. Rejoinder: The promotion of dogmatism. *Journal of Social Issues*, 1969, 25, 212-217. (d)
- Jensen, A. R. Counter Response. *Journal of Social Issues*, 1969, 25, 219-222. (e)
- Jensen, A. R. IQ's of identical twins reared apart. *Behavior Genetics*, 1970, 1, 133-148. (a)
- Jensen, A. R. Another look at culture-fair testing. In J. Hellmuth

- (Ed.), *Disadvantaged Child; Volume 3*. New York: Brunner/Mazel, 1970, 53-101. (b)
- Jensen, A. R. Can we and should we study race differences? In J. Hellmuth (Ed.), *Disadvantaged Child; Volume 3*. New York: Brunner/Mazel, 1970, 124-157. (c)
- Jensen, A. R. Selection of minority students in higher education. *The University of Toledo Law Review*, 1970, 1970, 403-457. (d)
- Jensen, A. R. Note on why genetic correlations are not squared. *Psychological Bulletin*, 1971, 75, 223-224.
- Jinks, J. L. and Fulker, D. W. Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychological Bulletin*, 1970, 73, 311-349.
- Kuttner, R. E. (Ed.), *Race and Modern Science*. New York: Social Science Press, 1967.
- Lane, E. A., Albee, G. W., and Doll, L. S. The intelligence of children of schizophrenics. *Developmental Psychology*, 1970, 2, 315-317.
- Lesser, G. S. Fifer, G., and Clark, D. H. Mental abilities of children from different social-class and cultural groups. *Monographs of the Society for Research in Child Development*, 1965, 30, No. 4, 1-115.
- Palmer, F. H. Socioeconomic status and intellectual performance among Negro pre-school boys. *Developmental Psychology*, 1970, 3, 1-9.
- Sattler, J. M. Racial 'experimenter effects' in experimentation, testing, interviewing, and psychotherapy. *Psychological Bulletin*, 1970, 73, 137-160.
- Scriven, M. The values of the academy (moral issues for American education and educational research arising from the Jensen case). *Review of Educational Research*, 1970, 40, 541-548.
- Shuey, A. M. *The Testing of Negro Intelligence*, Second edition, New York: Social Science Press, 1966.
- Stanley, J. C. Predicting college success of the educationally disadvantaged. *Science*, 1971, 171, 640-647.

ROBERT A. GORDON

The Johns Hopkins University

- Joseph A. Steger (Ed.). *Readings in Statistics for the Behavioral Sciences*. New York: Holt, Rinehart and Winston, 1971. Pp. ix + 406. \$5.95 (paperback).

This book of readings contains thirty-three articles divided into five chapters. The editor's stated purpose for the book is "... to supplement the basic courses in statistical methods and research design, or other undergraduate or first level graduate courses" (p. v). The designated audience is "... those who are not statisticians but who use statistics as tools in their field of study" (p. v).

Chapter one, entitled "Measurement and Statistics," is concerned with scales of measurement. A presentation by Stevens (1951) of his four scales of measurement is placed first, followed

by three additional articles which argue for less demanding prescriptions.

Chapter two deals entirely with Chi Square. It begins with the lengthy article by Lewis and Burke (1949) on "The Use and Misuse of the Chi Square Test," followed by five "comments," "replies," and "notes" defending or disputing the notions of Lewis and Burke. A fairly technical article by Cochran (1954) ends this chapter.

Chapter three is brief, dealing with "Parametric Techniques." It includes an historical article by "Student" (no reference) on the Lanarkshire milk experiment, an article dealing with transformations, and one dealing with analysis of covariance.

Chapter four is the book's lengthiest. Entitled "Assumptions and Statistical Inference," this chapter divides into three sub-sections. The first contains seven articles and notes on one-tailed vs. two-tailed hypothesis testing. The second focuses on null hypothesis testing, with six articles. The final sub-section contains two articles dealing with the effects of violations of assumptions in parametric tests, specifically, t and ANOVA.

Finally, Chapter five is entitled "Potpourri," and, as the title suggests, contains four unrelated articles. The most noteworthy occupant of this chapter is Walker's classical presentation of "Degrees of Freedom" (1940).

Because the editor makes a strong plea for use of the book as a supplementary text, it seems most reasonable to evaluate it primarily on pedagogical grounds. A brief section preceding the readings, "On Learning Statistics," is included apparently to enhance the book's value to a learner. This section attempts to convince students that in spite of their preconceptions, they can learn statistics. It also contains a list of commonly used symbols. This attempt, albeit brief, at making the book student-oriented is commendable. However, other pedagogical devices, such as end of chapter summaries, objectives, study questions, etc., were noticeably absent. On the positive side, a reasonably thorough index is included.

The editor wrote an introduction to each chapter, but they tended to be all too typical of books of readings, consisting essentially of summaries of the ensuing articles with little integration. In one noteworthy place, an "everyday" example was well used to introduce the readings (pp. 6-8). However, within this same section Stevens' four scales were defined. This was ground also covered by the first article and hence, probably not necessary.

Selecting articles for a book of readings in statistics with the purpose stated for this book is admittedly difficult. Articles from psychological journals (the major source for this book) are typically conceptually and/or mathematically difficult. Furthermore, often the material is quite specific and technical. Hence, finding articles

which are both of general interest and understandable to a beginning student would seem to be difficult, if not impossible.

Granting this difficulty, the reviewer feels the editor was not successful in achieving his goals. There seems to be few articles that a majority of beginning students could understand to the extent of making their reading worthwhile. Indeed, some articles included would seem to be obscure to almost anyone other than a professional statistician, e.g., "Analysis of Covariance: Its Nature and Uses" by Cochran (1957). The editor did state that "... the readings have been edited to abridge the highly technical material where possible" (p. v). However, as well as this reviewer could assess, the "where possible" translated to "seldom" and resultant editing did not alleviate the problem to any extent.

The editor included three controversial issues which proved not an unmixed blessing. Showing students that statisticians do not always agree, in fact, that they sometimes rival sophists with the energy they can devote to small metaphysical disputes, may be very worthwhile. Ritualized statistical practices are certainly far too prevalent. However, these controversies take up much of the book, leaving it ultimately unbalanced.

Seventy-seven pages are devoted to the use and misuse of the Chi Square statistic. The timeliness of this issue is reflected in the fact that the latest article included was published in 1950! Twenty-nine pages focus on one-tailed vs. two-tailed hypothesis testing (the latest article was published in 1954!) This reviewer personally found both of these sub-sections to be quite tedious.

Finally, eighty-three pages confront null hypothesis testing, beginning with Rozeboom's "The Fallacy of the Null-Hypothesis Test," (1960), and ending with what seemed to be the long-awaited *coup de grace*, David Bakan's first chapter in *On Method: Toward a Reconstruction of Psychological Investigation* (1967). Of the three controversies (four, if measurement scales is included), this was no doubt the best, being more contemporary and less tedious. However, beyond the somewhat casual mention of Bayesian statistics as an alternative, the student is left with hypothesis testing at his feet in a shambles, with nothing to take its place. This leads back to the lack of balance of the book.

To be sure, any editor must be selective in choosing subjects. Statistics is a vast subject. Nonetheless, to exclude all of regression and strength statistics (e.g., r , η^2 , ω^2) without as much as an introductory mention of their existence seems indefensible. One gets the impression by implication, both from the title and from the preface and introductions, that this book covers most of statistics used by behavioral scientists. In fact, hypothesis testing is emphasized at the expense of almost all else.

There is another count on which the title is misleading. The book

does not seem to be written for the behavioral scientist, but rather for the psychologist. This can be witnessed by the heavy use of material from psychological journals and books (26 out of the 31 cited articles¹). Moreover, it is highlighted by Dukes' article (" $N = 1$," 1965) which is completely concerned with psychological research.

In sum, the editor set out to put together a book of readings for "non-statistician" statistics students, from basic to graduate. The book does contain some excellent articles, one which students would do well to read. The inclusion of several controversies also has merit. Nonetheless, for reasons given above, this reviewer feels that the book has limited pedagogical value. The basic idea of a supplementary text for a given subject which contains writings of people "in the field" seems sound. Whether this can be successfully accomplished in statistics is not answered by this book.

REFERENCES

- Bakan, D. The test of significance in psychological research. *On method: Toward a reconstruction of psychological investigation*. San Francisco: Jossey-Bass, Inc., 1967, pp. 1-29.
- Cochran, W. G. Analysis of covariance: Its Nature and Uses. *Biometrics*, 13, 1957, 261-281.
- Cochran, W. G. Some methods for strengthening the common tests. *Biometrics*, 10, 1954, 417-451.
- Dukes, W. F. " $N = 1$." *Psychological Bulletin*, 64, 1965, 74-79.
- Lewis, D. and Burke, C. J. The use and misuse of the chi-square test. *Psychological Bulletin*, 46, 1949, 433-489.
- Rozeboom, W. W. The fallacy of the Null-Hypothesis Significance Test. *Psychological Bulletin*, 57, 1960, 416-428.
- Stevens, S. S. *Handbook of experimental psychology*. New York: John Wiley and Sons, 1951, pp. 23-30.
- Walker, H. M. Degrees of Freedom. *Journal of Educational Psychology*, 31, 1940, 253-269.

GERALD M. GILLMORE
University of Illinois
(Urbana-Champaign Campus)

¹ Neither the reference of an article by Bartlett, "The Use of Transformations," nor the reference of the article by Student (mentioned earlier) was cited. These are very unfortunate printing errors.

EDUCATIONAL and
PSYCHOLOGICAL



MEASUREMENT

Editor: W. Scott Gehman

Managing Editor: Geraldine R. Thomas

BOARD OF COOPERATING EDITORS

- DOROTHY C. ADKINS, *University of Hawaii*
LEWIS R. AIKEN, JR., *Guilford College*
HAROLD P. BECHTOLDT, *The University of Iowa*
WILLIAM V. CLEMANS, *Science Research Associates, Inc.*
LOUIS D. COHEN, *University of Florida*
JUNIUS A. DAVIS, *Educational Testing Service*
HAROLD A. EDGERTON, *Performance Research, Inc.*
MAX D. ENGELHART, *Duke University*
GENE V. GLASS, *University of Colorado*
E. B. GREENE, *Chrysler Corporation (Retired)*
J. P. GUILFORD, *University of Southern California, Los Angeles*
JOHN A. HORNADAY, *Babson College*
JOHN E. HORROCKS, *The Ohio State University*
CYRIL J. HOYT, *University of Minnesota*
MILTON D. JACOBSON, *University of Virginia*
JOSEPH C. JOHNSON II, *Duke University*
WILLIAM G. KATZENMEYER, *Duke University*
E. F. LINDQUIST, *State University of Iowa*
FREDERIC M. LORD, *Educational Testing Service*
ARDIE LUBIN, *Naval Medical Neuropsychiatric Research Unit, San Diego*
LOUIS L. MCQUITT, *University of Miami, Coral Gables*
WILLIAM B. MICHAEL, *University of Southern California, Los Angeles*
HOWARD G. MILLER, *North Carolina State University at Raleigh*
ELLIS B. PAGE, *The University of Connecticut*
NAMBURY S. RAJU, *Science Research Associates, Inc.*
BEN H. ROMINE, JR., *University of North Carolina at Charlotte*
KENDON SMITH, *The University of North Carolina at Greensboro*
THELMA G. THURSTONE, *University of North Carolina at Chapel Hill*
HERBERT A. TOOPS, *The Ohio State University*
WILLARD G. WARRINGTON, *Michigan State University*
JOHN E. WILLIAMS, *Wake Forest University*
E. G. WILLIAMSON, *University of Minnesota*

COMPUTER PROGRAMS

This section is provided for the early publication at the expense of the author of computer programs relevant to measurement in the fields of education and psychology. Customarily, a program should be expected not to exceed six or eight printed pages. Manuscripts of four or fewer printed pages are preferred. Each manuscript will be carefully reviewed as to its suitability and accuracy of content. In some instances an accepted paper may be returned to the author for possible revisions or shortening. The cost to the author will be forty-five dollars per page plus ten dollars extra per page for tables, figures, and formulas.

Authors are granted permission to have reprints made of their articles at their own expense.

Manuscripts received up to September first will be considered for the Spring issue; manuscripts received between then and March first will be considered for the Autumn issue.

All correspondence and duplicate manuscripts should be directed to:

Dr. William B. Michael
325 Callita Place
San Marino, California 91108.

A THEORETICAL STUDY OF THE MEASUREMENT EFFECTIVENESS OF FLEXILEVEL TESTS¹

FREDERIC M. LORD

Educational Testing Service

A conventional test becomes a flexilevel test when modified so that the examinee follows these rules:

1. Answer first a specified test item of median difficulty.
2. After answering an item correctly, attempt next the easiest unanswered item of more-than-median difficulty. After answering an item incorrectly, attempt next the hardest unanswered item of less-than-median difficulty.

A special answer sheet is used so that the examinee will know whether each answer is correct or incorrect. If the conventional test contains N items, the examinee taking the flexilevel test will attempt only $n = (N + 1)/2$ of these. A method for implementing flexilevel testing is described by Lord (1971b).

Surprisingly, it appears that number-right scoring is quite effective for flexilevel tests (Lord, 1971b), in spite of the fact that different examinees answer different sets of items. A worthwhile refinement, used throughout the research reported here, is to add one-half score point to the number-right score of each examinee who answered his last-attempted item incorrectly.

A crucial question is whether flexilevel testing will be too confusing or too time-consuming for many examinees. Empirical studies

¹ This research was sponsored in part by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under Contract No. N00014-69-C-0017, Contract Authority Identification Number, NR No. 150-303, and Educational Testing Service. Reproduction in whole or in part is permitted for any purpose of the United States Government.

are needed to answer this and other questions of practical effectiveness.

Since a theoretical study can be done more quickly and less expensively than a substantial empirical study, the study reported here was carried out in order to evaluate various flexilevel tests from a theoretical point of view. An important purpose was to try to separate some flexilevel designs that are worth trying out empirically from those that are altogether inferior to other tests.

In order to carry out a theoretical investigation of this type, it is necessary to be able to predict probabilistically how a given examinee will respond to items different from those already administered. Consequently, the present results are derived from item characteristic curve theory (see, for example, Lord, 1970, sections 3-4).

Here we assume the probability P_i that a given examinee will answer item i correctly depends *only* on his "ability" level, denoted by θ , and on certain item parameters: a ("discriminating power"), b ("difficulty"), and c ("pseudo chance-score level"). These item parameters are assumed to have been already determined, to an adequate approximation, by pretesting.

Conditional Frequency Distribution of Test Score

We can evaluate any given flexilevel test once we can determine $f(x | \theta)$, the conditional frequency distribution of test scores x for examinees at ability level θ . Given some mathematical form for the function $P_i \equiv P_i(\theta) \equiv P(\theta; a_i, b_i, c_i)$, the value of $f(x | \theta)$ can be determined numerically for any specified value of θ by the recursive method outlined below.

Assume the N test items to be arranged in order of difficulty, as measured by the parameter b_i . We will choose N to be an odd number. For present purposes (not for actual test administration) identify the items by the index i , taking on the values $-n + 1, -n + 2, \dots, -1, 0, 1, \dots, n - 2, n - 1$, respectively, when the items are arranged in order of difficulty. Thus b_0 is the median item difficulty.

Consider, for example, the sequence of right (R) and wrong (W) answers

$R \quad W \quad W \quad R \quad W \quad R \quad R \quad R \quad W \quad R.$

Following the rules given for a flexilevel test, we see that the corresponding sequence of items answered is

$i = 0, +1, -1, -2, +2, -3, +3, +4, +5, -4, +6.$

The general rule is that if item i is the v th item administered and item j is the $(v + 1)$ th, then, for flexilevel tests,

either $j = i + 1$ or $j = i - v$ when $i \geq 0$,

either $j = i - 1$ or $j = i + v$ when $i \leq 0$.

In the same context, let $P_{ij,v} = P_{ij,v}(\theta)$ denote the probability that item j will be the next item administered after item i .

$$\text{If } i \geq 0, \quad P_{i,j,v} = \begin{cases} P_i(\theta) & \text{if } j = i + 1, \\ Q_i(\theta) & \text{if } j = i - v, \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{If } i \leq 0, \quad P_{i,j,v} = \begin{cases} P_i(\theta) & \text{if } j = i + v, \\ Q_i(\theta) & \text{if } j = i - 1, \\ 0 & \text{otherwise.} \end{cases}$$

For examinees at ability level θ , let $p_v(i | \theta)$ denote the probability that item i is the v th item administered. Clearly,

$$p_{v+1}(j | \theta) = \sum_{i=-n+1}^{n-1} p_v(i | \theta) P_{i,j,v}(\theta). \quad (1)$$

Now, the first item administered ($v = 1$) is always item $i = 0$, so

$$p_1(i | \theta) = \begin{cases} 1 & \text{if } i = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Starting with this fact and with a knowledge of all the $P_i(\theta)$ (determined from pretest data), equation (1) allows us to compute the values of $p_v(i | \theta)$ for each i , for $v = 2, 3, \dots, n$, and for any specified set of values of θ .

Now we can make use of a readily verified feature of flexilevel tests. Again let j represent the $(v + 1)$ th item to be administered. If $j > 0$, then the number-right score r on the v items already administered is $r = j$; if $j < 0$, then $r = v + j$.

Thus the frequency distribution of the number-right score r for examinees at ability level θ is given by $p_{n+1}(r | \theta)$ for those examinees who answered correctly the n th (last) item administered, by $p_{n+1}(r - n | \theta)$ for those who answered incorrectly. This frequency distribution can be computed recursively from (1).

As already noted, the actual score assigned on a flexilevel test is $x = r$ if the last item is answered correctly, $x = r + \frac{1}{2}$ if it is answered incorrectly. Consequently the conditional distribution of test scores is

$$f(x | \theta) = \begin{cases} p_{n+1}(x | \theta) & \text{if } x \text{ is an integer,} \\ p_{n+1}(x - n - \frac{1}{2} | \theta) & \text{if } x \text{ is a half integer.} \end{cases} \quad (2)$$

For any specified test design, this conditional frequency distribution $f(x | \theta)$ can be computed for $x = 0, \frac{1}{2}, 1, 1\frac{1}{2}, \dots, n$ for various values of θ . Such distributions constitute the totality of possible information relevant to evaluating the effectiveness of x as a measure of ability.

Evaluating a Flexilevel Testing Procedure

If we are to use x as a measure of ability, we would like $\mu_x | \theta_1$ (the mean of x when $\theta = \theta_1$) to differ from $\mu_x | \theta_2$ whenever $\theta_1 \neq \theta_2$. It seems natural to use the "critical ratio"

$$\frac{\mu_x | \theta_2 - \mu_x | \theta_1}{\sigma_x | \theta}$$

to summarize the effectiveness of x for discriminating between ability levels θ_1 and $\theta_2 = \theta_1 + \Delta$, where $\sigma_x | \theta$ is the conditional standard deviation of x and Δ represents a small increment in ability (small enough so that $\sigma_x | \theta = \sigma_x | \theta + \Delta$ approximately).

Actually we will work with the square of this ratio:

$$I_x(\theta) = \frac{k(\mu_x | \theta + \Delta - \mu_x | \theta)^2}{\sigma_x | \theta^2}, \quad (3)$$

where k is any convenient constant. Given some small increment Δ , $I_x(\theta)$, as a function of θ , is readily computed from (2) for any specified test design. Since we are only interested in comparisons between designs, the values of k and Δ are of no importance so long as they are the same for all designs compared.

Test Designs Studied

The numerical results reported here are obtained on the assumption that P_i is a normal ogive, possibly modified to accommodate the effects of success due to guessing:

$$P_i = P(\theta; a, b_i, c) = c + (1 - c) \int_{-\infty}^{a(\theta - b_i)} \phi(t) dt, \quad (4)$$

where $\phi(t)$ is the normal density function. The results would presumably be about the same if P_i had been assumed logistic rather than normal ogive.

To keep matters simple, we will only consider tests in which all items have the same discriminating power, a ; also the same pseudo chance level, c . Results are presented here separately for $c = 0$ (no guessing) and $c = .2$. The results are general for any value of $a > 0$, since a can be absorbed into the unit of measurement chosen for the ability scale (as will be noticed for the base line shown in the figures).

In all tests studied, each examinee answers exactly $n = 60$ items. For simplicity, we will consider only tests in which the item difficulties form an arithmetic sequence, so that $b_{i+1} - b_i = d$, say.

Results for Tests with No Guessing

Figure 1 compares the effectiveness of four 60-item ($n = 60$, $N = 119$) flexilevel tests with each other and with three benchmark tests. The scale chosen for θ in the figures is such that for typical achievement and aptitude tests the standard deviation of θ in typical high school and college groups will be very roughly $\sigma_\theta = 1/2a$ (a more detailed explanation is given in Lord (1971a)).

The "standard test" is a conventional 60-item test composed entirely of items of difficulty $b = 0$, scored by counting the number of right answers. There is no guessing, so $c = 0$. The values of a and c are the same for benchmark and flexilevel tests. For fixed a and c , no test composed of dichotomously scored items with characteristic curves (4) can have a higher value of $I_x(\theta)$ at any θ than the standard test has at $\theta = b_0$ (see Birnbaum, 1968).

As would be expected, the figure shows that the standard test is best for discriminating among examinees at ability levels near $\theta = 0$. If good discrimination is important at $\theta = \pm 2/2a$ or $\theta = \pm 3/2a$, then a flexilevel test such as the one with $d = .033/2a$ or $d = .050/2a$ is better. The larger d is, the poorer the measurement at $\theta = b_0$, but the better the measurement at extreme values of θ .

Suppose the best possible measurement is required at $\theta = \pm 2$, with $a = 0.5$. It might be thought that an effective conventional 60-item test for this limited purpose would consist of 30 items at $b = +2$ and 30 items at $b = -2$. The curve for this last test is shown in Figure 1. The fact is that with $a = 0.5$, no unpeaked test

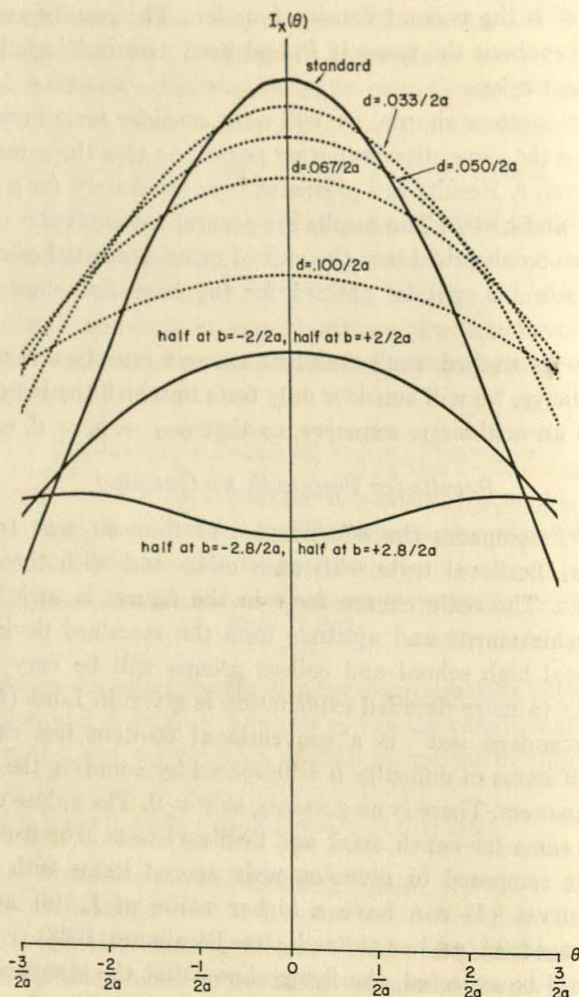


Figure 1. Relative efficiency of four 60-item flexilevel tests with $b_0 = 0$ (curves with d s) and three bench mark tests. $c = 0$.

(i.e., no test with items at more than one difficulty level) can simultaneously measure as well as both $\theta = +2$ and $\theta = -2$ as does the standard test (which has all items peaked at $b = 0$).

The situation is different if the best possible measurement is required at $\theta = \pm 3$, with $a = 0.5$. Using dichotomously scored items, the best 60-item conventional test for this purpose consists of 30

items at $b = -2.8$ and 30 items at $b = +2.8$, approximately. The curve for this test is shown in Figure 1.

For fixed θ , the number-right score x on a standard test has a binomial distribution. Thus, the expected score is

$$\mu_{x|\theta} = nP$$

and the variance of the scores is

$$\sigma_{x|\theta}^2 = nPQ,$$

where $P = P(\theta)$ is given by (4). It is apparent from (3) that $I_x(\theta)$ for a standard test is proportional to n , the test length.

We now see that when $a = 0.5$, the 60-item flexilevel test with $d = .033$ gives about as effective measurement as a

- 58-item standard test at $\theta = 0$,
- 60-item standard test at $\theta = \pm 1$,
- 69-item standard test at $\theta = \pm 2$,
- 86-item standard test at $\theta = \pm 3$.

At $\theta = \pm 3$, the 60-item flexilevel test with $d = .1$, is as effective as a 96-item standard test.

Results for Tests with Guessing

Figure 2 compares the effectiveness of three 60-item flexilevel tests with each other and with five bench mark tests. All items have $c = 0.2$ and all have the same discriminating power a . The standard test is a conventional 60-item test with all items at difficulty level $b = 0.5/2a$, scored by counting the number of right answers.

If all the item difficulties in any test were changed by some constant amount Δb , the effect would be simply to translate the corresponding curve by an amount Δb along the θ -axis. The difficulty level of each bench mark test and the starting item difficulty level b_0 of each flexilevel test in Figure 2 has been chosen so as to give maximum discriminating power somewhere in the neighborhood of $\theta = 0$.

The standard test is again found to be best for discriminating among examinees at ability levels near $\theta = 0$. At $\theta = \pm 2$ the flexilevel tests are better than the standard test, which in turn seems to be better than any of the other conventional (bench mark) tests, although the situation is less clear than before because of the asymmetry of the curves.

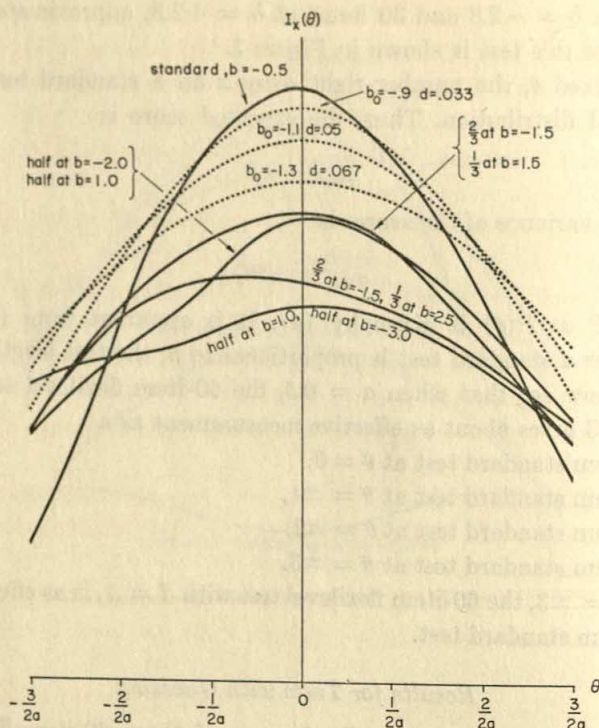


Figure 2. Relative efficiency of three 60-item flexilevel tests (curves with ds) and five bench mark tests. $c = 0.2$. (Numerical labels on curves are for $a = 0.5$.)

When $a = 0.5$ the 60-item flexilevel test with $b_0 = -0.9$ and $d = .033$ gives about as effective measurement as a

58-item standard test at $\theta = 0$

60-item standard test at $\theta = \pm 1$

70-item standard test at $\theta = -2.0$ or $\theta = +2.25$

83-item standard test at $\theta = +3$

114-item standard test at $\theta = -3$

at $\theta = -3$, the 60-item flexilevel test with $b_0 = -1.3$ and $d = .067$ is as effective as a 137-item standard test.

Conclusion

Near the middle of the ability range for which the test is designed, a flexilevel test is less effective than is a comparable peaked conventional test. In the outlying half of the ability range, the flexilevel

test provides more accurate measurement in typical aptitude and achievement testing situations than a peaked conventional test composed of comparable items. This comparison assumes that 60 items are administered to each examinee. The advantage of flexilevel tests over conventional tests at low ability levels is significantly greater when there is guessing than when there is not.

Empirical studies will be needed to answer such questions as the following:

1. To what extent are different types of examinees confused by flexilevel testing?
2. To what extent does flexilevel testing lose efficiency because of an increase in testing time per item?
3. How adequately can we score the examinee who does not have time to finish the test?
4. How can we score the examinee who does not follow directions?
5. What other serious inconveniences and complications are there in flexilevel testing?
6. Is the examinee's attitude and performance improved when a flexilevel test "tailors" the test difficulty level to match his ability level?

Empirical investigations should study tests designed in accordance with the theory used here. Otherwise, it is likely that a poor choice of d and especially of b_0 will result in an ineffective measuring instrument.

The most likely application of flexilevel tests is in situations where it would otherwise be necessary to unpeak a conventional test in an attempt to obtain adequate measurement at the extremes of the ability range. Such situations are found in nationwide college admissions testing and elsewhere.

REFERENCES

- Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968. Chapters 17-20.
- Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer assisted instruction, testing, and guidance*. New York: Harper and Row, 1970. Chapter 8.
- Lord, F. M. A theoretical study of two-stage testing. *Psychometrika*, 1971, 36, 227-242. (a)
- Lord, F. M. The self-scoring flexilevel test. *Journal of Educational Measurement*, 1971, 8, 147-151. (b)

A SHORT CUT TOWARD A SUBMATRIX CONTAINING ONLY "DISTURBED" INDIVIDUALS^{1, 2}

LOUIS L. McQUITTY

University of Miami, Coral Gables

Two recent studies found "normal" individuals relatively like one another and "disturbed" individuals relatively unlike one another, with intermediate degrees of association between members from each of these two categories.

The above differences express themselves most clearly when the comparisons are restricted to the relatively large and relatively small indices between individuals. A computer program was developed for comparing every submatrix with every other submatrix and a criterion was developed, applied, and found helpful for determining which pair would most likely yield the best separation of "normal" and "disturbed" individuals (McQuitty, Banks, and Frary, 1970; McQuitty, Banks, Frary, and Aye, 1972).

Hypothesis

The above approach is so elaborate that it can analyze only small matrices (approximately 16×16). A simple way is here developed and illustrated for continuing to divide and redivide a matrix of inter associations between "normal" and "disturbed" subjects into submatrices which are hypothesized to contain only "normals" in one submatrix and only "disturbed" subjects in the other submatrix.

¹ This investigation was supported by Public Health Service Research Grant No. MH 14070-03 from National Institute of Mental Health.

² Appreciation is expressed to Elizabeth Ann McQuitty, who performed all of the calculations by pencil and paper on the small sample of data (except the agreement scores of the original matrix). The calculations were later confirmed by the development of a computer program.

*First Study—Small Sample**The Data*

The method is illustrated with the data of Table 1 which reports agreement scores between eight "normal" and eight "disturbed" subjects.

Clinicians selected both categories of subjects from among those seeking counseling from the Counseling Center at Michigan State University during the academic year of 1966-67. The "disturbed" subjects were chosen as in need of psychotherapy and the "normal" subjects as not in need of psychotherapy.

Each subject completed a test illustrated by the following two items: "The word mother suggests hope. "yes no ?" and "The word father suggests hate. yes no ?."

There were 13 concepts (1. control, 2. self, 3. marriage, 4. religion, 5. father, 6. achievement, 7. woman, 8. closeness, 9. distance, 10. dependency, 11. sex, 12. man, and 13. friend) and ten emotions (1. fear, 2. loneliness, 3. love, 4. hate, 5. guilt, 6. hope, 7. anxiety, 8. anger, 9. frustration, and 10. depression). Every concept was associated with every emotion as illustrated in the above two items to yield a test of 130 items.

Any two subjects have an agreement on an item if and only if both answer "yes," "no," or "?" on the item. Their total agreement score is the number of items on which they agree. The agreement scores of every subject with every other subject are reported in the matrix of Table 1.

The Method

The two largest and the two smallest entries of every column of Table 1 are underlined, summed, and reported in Row *a* at the bottom of the table. (Even in the case of a tie, only two of the larger and only two of the smaller scores are underlined.) In the general case about one-fourth to one-third of the entries in each column are underlined, one-half of them the larger entries and the other one-half of them the smaller entries.

The sums of Row *a* are ranked in Row *b*, with the largest sum assigned a rank of one.

If *d* (the number of "disturbed" subjects) is greater than or approximately equal to *n* (the number of "normals"), one-half of

TABLE 1
A Matrix of Agreement Scores between "Normal" and "Disturbed" Subjects

Classification Code Numbers	N	1	N	2	D	3	D	4	D	5	D	N	6	D	7	N	8	D	9	D	10	N	11	D	12	N	13	N	14	N	15	D	16
1																																	
2																																	
3																																	
4																																	
5																																	
6																																	
7																																	
8																																	
9																																	
10																																	
11																																	
12																																	
13																																	
14																																	
15																																	
16																																	
Row a																																	
Row b																																	

N = "Normal"; D = "Disturbed"; Row a = sum of underlined entries; Row b = Ranks of entries of Row a.

d subjects, those with the larger numerical ranks (smaller sums), are withdrawn from the matrix and hypothesized to be "disturbed" subjects. In the present case, these are Subjects 3, 4, 12, and 16, with ranks of 16, 13, 14, and 15 respectively. All of them are in fact "disturbed" subjects. If the n is considerably greater than d , one-half of the "normal" subjects are withdrawn first (as illustrated in the next step of this example).

The remaining subjects are entered in a new matrix, in this case a 12×12 matrix, as shown in Table 2, using the agreement scores from Table 1.

The same steps as applied above to Table 1 are now applied to Table 2, except that when the number of "normal" subjects is greater than the number of "disturbed" subjects, approximately one-half of the n subjects are withdrawn and predicted to be "normal." The subjects with the smaller numerical ranks (larger agreement scores) are withdrawn. In this case they are Subjects 1, 2, 9, 13, and 11, with ranks of 4.5, 2.0, 3.0, 1.0, and 4.5. They are all in fact "normal" except Subject 9 who is "disturbed."

Five subjects, rather than four (one-half of the n subjects) are withdrawn and predicted to be "normal" because the fourth and fifth subjects are tied with a rank of 4.5.

The remaining subjects are assembled in a new matrix, Table 3. In this case, there are seven subjects. The same steps as outlined for Table 1 are applied to Table 3, except that d' subjects (the number of "disturbed" subjects still predicted to be in the matrix) are withdrawn and predicted to be "disturbed." Only the sums of the largest and smallest scores in each column are used in the predictions because these two scores represent one-third of the entries in each column. Since four of a total of eight subjects had already been predicted to be "disturbed," this left d' equal to four. The four subjects with the higher numerical ranks (smallest sums of agreement scores) are withdrawn and predicted to be "disturbed." They are Subjects 10, 5, 7, and 15, with ranks of 4, 5, 6.5, and 6.5, respectively. All but Subject 15 are in fact "disturbed." The remaining three subjects are predicted to be "normal." All of them are in fact "normal."

Results

Of the eight subjects predicted to be "disturbed," seven of them

TABLE 2
Agreement Scores of Table 1 without those Subjects Predicted by Table 1 to be "Disturbed"

Classification Code Numbers	N	1	N	2	N	6	N	8	D	9	D	10	N	13	N	14	D	5	D	7	N	11	N	15
1		<u>101</u>		101		86		95		95		<u>101</u>		100		82		84		81		88		86
2		<u>101</u>				<u>83</u>		84		100		94		<u>115</u>		88		<u>93</u>		86		<u>106</u>		84
6		86		<u>83</u>				90		93		87		94		84		84		87		<u>78</u>		<u>92</u>
8		95		<u>84</u>		90				<u>104</u>		96		93		<u>98</u>		93		84		<u>77</u>		79
9		95		100		<u>93</u>		<u>104</u>				<u>98</u>		99		93		<u>101</u>		<u>97</u>		91		80
10		<u>101</u>		94		87		96		98				91		86		86		<u>76</u>		83		78
13		100		<u>115</u>		<u>94</u>		93		99		91		94		<u>94</u>		93		89		<u>104</u>		<u>90</u>
14		82		88		84		<u>98</u>		93		86		93		92		92		90		80		79
5		84		93		84		93		<u>101</u>		86		93		92		86		86		83		73
7		81		86		87		84		<u>97</u>		<u>76</u>		<u>89</u>		90		86				80		75
11		88		<u>106</u>		<u>78</u>		<u>77</u>		<u>91</u>		83		<u>104</u>		<u>80</u>		<u>83</u>		80		84		84
15		86		84		92		<u>79</u>		80		<u>78</u>		<u>90</u>		<u>351</u>		<u>73</u>		<u>75</u>		365		330
Row a		365		388		348		358		376		353		398		351		350		338		365		330
Row b		4.5		2		10		6		3		7		1		8		9		11		4.5		12

N = "Normal"; *D* = "Disturbed"; Row a = sum of underlined entries; Row b = ranks of entries of Row a.

TABLE 3

Agreement Scores of Table 2 without those Subjects Predicted by Table 2 to be "Normal"

Classification Code Number	N 6	N 8	D 10	N 14	D 5	D 7	N 15
6		90	87	84	84	87	<u>92</u>
8	90		<u>96</u>	<u>98</u>	<u>93</u>	84	79
10	87	96		86	86	76	78
14	<u>84</u>	<u>98</u>	86		92	<u>90</u>	79
5	84	93	86	92		86	<u>73</u>
7	87	84	<u>76</u>	90	86		75
15	<u>92</u>	<u>79</u>	78	<u>79</u>	<u>73</u>	<u>75</u>	
Row a	176	177	172	177	166	165	165
Row b	3	1.5	4	1.5	5	6.5	6.5

N = "Normal"; D = "Disturbed"; Row a = sum of underlined entries; Row b = ranks of entries of Row a.

are "disturbed" as evaluated by clinicians, and of the eight subjects predicted to be "normal," seven of them are "normal" as evaluated by clinicians. These results are identical, subject for subject, with those obtained by the more elaborate method (McQuitty, Banks, Frary, and Aye, 1972). Eighty-eight per cent of the subjects were correctly classified to yield a phi coefficient of 0.75 and a chi-square of 9.00 with a significance of .0016 on a one-tailed test.

Interpretation

The method of this paper is a rapid short cut to the elaborate process of comparing every submatrix of d subjects (the number of "disturbed" subjects in a matrix) with the other submatrix of n subjects (the number of "normal" subjects in the matrix) to obtain that pair of submatrices which gives the best differentiation between the members of the two categories in terms of some objective criterion.

The method was developed in relation to a set of data which had been carefully studied by the investigator. As a result the method may have features adapted to unique characteristics of that data. Its apparent success with that data might depend in part on "chance." The method is, therefore, applied below to a larger and independent set of data.

Second Study—Large Sample

Theory

The theory maintains that psychological disturbance expresses

itself in emotional components of intra-individual concepts. Examples of intra-individual concepts are: (1) my behavior, (2) my heart, (3) my reputation, and (4) my attitude. Each such concept is presumed to have emotional flavors which differ from individual to individual. Psychological disturbance is presumed to express itself in many kinds of interrelationships among these emotional flavors and often in restricted and unique patterns. Psychologically disturbed individuals are thought to relate uniquely to both "normals" and other "disturbed" individuals in these emotional flavors.

The Test

The above emotional flavors can presumably be tapped by such test items as the following ones:

- | | | | |
|----------------------------------|-----|----|---|
| 1. My behavior suggests hope. | yes | no | ? |
| 2. My heart suggests love. | yes | no | ? |
| 3. My reputation suggests guilt. | yes | no | ? |
| 4. My attitude suggests sadness. | yes | no | ? |

The test derived from 14 intra-individual concepts and 14 emotions, every emotion associated with every concept in the fashion illustrated above to yield 196 items. The emotions and concepts are as follows: (1) hope, (2) love, (3) guilt, (4) joy, (5) sadness, (6) hate, (7) sympathy, (8) fear, (9) anger, (10) pride, (11) anxiety, (12) happiness, (13) respect, (14) distrust, and (1) my behavior, (2) my heart, (3) my reputation, (4) my attitude, (5) my soul, (6) my past, (7) my beliefs, (8) my conscience, (9) my religion, (10) myself, (11) my future, (12) my state of mind, (13) my feelings, and (14) my body. No statistical analysis has yet been applied in an effort to improve the test.

The Subjects

The subjects were 144 undergraduate students from the University of Miami, Coral Gables, Florida, who sought counseling voluntarily at the University Counseling Center between September 13 and December 20, 1969. The test was administered before a screening interview for the purpose of gathering routine information and assigning the subject to a clinical or vocational counselor.

After having been seen at least twice by a counselor, every subject was classified by his counselor as "disturbed" or "nondis-

turbed." In this context a "disturbed" subject was one who, in the opinion of the counselor, was experiencing a serious behavioral problem as a result of mental disturbance. Such problems ranged from those associated with the more acute forms of mental illness to neurotic reactions which inhibited academic performance or social interaction in a serious manner. A subject was not classified as "disturbed" if his problem was not judged to have a substantial effect on his overall performance as a student or his relationships with others.

As a result of the above approach, 66 of the subjects were classified by counselors as "disturbed" and 78 were classified as "nondisturbed" (i.e., "normal"). Every subject was classified by only one of six counselors. Table 4 shows the classifications of the subjects by the counselors.

The Analysis

The agreement score was computed between every subject with every other subject to yield a 144×144 matrix. It was analyzed seven times by the method described above, each time with the statistical decision based on a different number of entries in each column. In every case the decision was based on an equal number of higher and lower entries in each column. The first time it was based on four entries (the first and second largest entries plus the first and second smallest entries), the second time on one-fourth of the entries (the larger one-eighth plus the smaller one-eighth), etc., for one-third, one-half, two-thirds, three-fourths, and all entries.

TABLE 4
Classification of Subjects by Counselors

	Counselor Code	Classifications	
		N	D
1	A	8	12
2	B	25	8
3	C	13	15
4	D	14	8
5	E	13	22
6	F	5	1
Totals		78	66

N = "Nondisturbed" subjects; D = "Disturbed" subjects.

Results

Table 5 reports the percentages of correct selections when (a) one-half of the number of "disturbed" (33) were selected, (b) one-half of the number of "normals" (39) were selected, (c) 66 were selected as "disturbed," and (d) 78 were selected as "normal" together with the phi coefficients, chi squares, and level of significance for the chi square using a one-tailed test. The latter category of statistics is reported twice, once when half of the selections had been completed and again at the end of the selections.

The results obtained with one-fourth of the entries of every column are representative of or only slightly superior to those obtained when other numbers of entries were used. These results are summarized. The first two selections (which sought one-half of both the "disturbed" and the "normals") produced 78.8% and 79.5%, respectively, of correct classification as compared with 71.2% and 75.6%, respectively, for all subjects of these two categories and 73.6% for the two categories combined; a phi coefficient of .582 for one-half as compared with one of .469 for all subjects; and corresponding levels of significance for chi square of .000020 and .0000038.

Interpretation

The results are unusually encouraging; equal degrees of differentiation based on objective analysis of data from unimproved tests responded to by two highly similar categories of subjects have infrequently, if ever before, been achieved in this area. The method is of such a nature that it is not directed to taking advantage of chance errors. Cross validation, for the usual reasons, is not required. However, because of the unusual degree of differentiation, follow-up studies are desirable; very atypical results are possible by chance alone.

The findings show that certain "disturbed" individuals are in general relatively unlike themselves and "normals" to such an extent on certain psychological tests that this fact can be used to differentiate between "normal" and "disturbed" subjects in unusually high agreement with clinicians.

Two other methods based on the same or similar hypothesized relationships have been described elsewhere using the concepts of "spots" (McQuitty and Frary, 1971) and "scoring matrices."³

TABLE 5
Differentiation and Significance

No. of Entries*	Percentage Correct				Phi Coefficient		Chi square		Level of Significance One-tailed Test	
	33D's	39N's	66D's	78N's	66D's+ 78N's	33D's+ 39N's	66D's+ 78N's	33D's+ 39N's	66D's+ 78N's	33D's+ 39N's
4	66.7	73.2	68.2	73.1	70.8	.398	.413	10.19	22.88	.000038
1/4	78.8	79.5	71.2	75.6	73.6	.582	.469	22.07	29.75	.0000038
1/3	78.8	80.0	71.2	75.6	73.6	.587	.469	22.81	29.75	.0000038
1/2	78.8	79.5	71.2	75.6	73.6	.582	.469	22.07	29.75	.0000038
2/3	78.8	79.5	71.2	75.6	73.6	.582	.469	22.07	29.75	.0000038
3/4	78.8	79.5	71.2	75.6	73.6	.582	.469	22.07	29.75	.0000038
all	78.8	79.5	71.2	75.6	73.6	.582	.469	22.07	29.75	.0000038

* Number of entries in statistical decisions.
D = "Disturbed" subjects; N = "Nondisturbed" subjects.

Summary

If a matrix of interassociations between individuals is known to contain d "disturbed" and n "normal" subjects, it can be divided into all possible pairs of submatrices of size d and n and a criterion can be applied to determine the pair yielding the best classification of "normal" and "disturbed" subjects (McQuitty, Banks, Frary, and Aye, 1972). The present paper develops and illustrates a method many times shorter and found it to be more effective with an unimproved test than any other objective method reported (except the above mentioned long method) in the analysis of psychological tests for differentiating "disturbed" from "normal" college students.

REFERENCES

- McQuitty, L. L. and Frary, J. M. Reliable and valid hierarchical classification. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1971, 31, 321-346.
- McQuitty, L. L., Banks, R. G., and Frary, J. M. Submatrices of interassociations for scoring interrelatedness within matrices as an index of psychological disturbance. *Multivariate Behavioral Research*, 1970, 5, 479-488.
- McQuitty, L. L., Banks, R. G., Frary, J. M., and Aye, C. D. Selecting a submatrix likely to contain only "disturbed" subjects. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1972, in press.

It is a common knowledge that the medical profession is a "closed shop" and a "trust" and it is not to be wondered at that the public has a right to know the facts of the situation. The American Medical Association is a body of men who are interested in the health of the people and who are willing to do whatever is necessary to protect the public interest. The American Medical Association is a body of men who are interested in the health of the people and who are willing to do whatever is necessary to protect the public interest. The American Medical Association is a body of men who are interested in the health of the people and who are willing to do whatever is necessary to protect the public interest.

It is a common knowledge that the medical profession is a "closed shop" and a "trust" and it is not to be wondered at that the public has a right to know the facts of the situation. The American Medical Association is a body of men who are interested in the health of the people and who are willing to do whatever is necessary to protect the public interest. The American Medical Association is a body of men who are interested in the health of the people and who are willing to do whatever is necessary to protect the public interest. The American Medical Association is a body of men who are interested in the health of the people and who are willing to do whatever is necessary to protect the public interest.

It is a common knowledge that the medical profession is a "closed shop" and a "trust" and it is not to be wondered at that the public has a right to know the facts of the situation. The American Medical Association is a body of men who are interested in the health of the people and who are willing to do whatever is necessary to protect the public interest. The American Medical Association is a body of men who are interested in the health of the people and who are willing to do whatever is necessary to protect the public interest. The American Medical Association is a body of men who are interested in the health of the people and who are willing to do whatever is necessary to protect the public interest.

RELIABILITY OF MULTIPLE-CHOICE TESTS IS THE PROPORTION OF VARIANCE WHICH IS TRUE VARIANCE

EDWARD E. CURETON

University of Tennessee

IN a recent issue of this journal, Frary (1969) presents an analysis which seems to show that classical weak true-score theory does not apply to multiple-choice tests. Starting with the equation,

$$x = t + g + e, \quad (1)$$

he shows that the variance ratios $\sigma_{t+g}^2/\sigma_x^2$ and σ_t^2/σ_x^2 , the correlation between equivalent forms, and the square of the correlation between raw scores and true scores are all different.

The difficulty with this derivation is that the guessing score g is not separated into a true component and an error component.

Guessing tendency is a real trait on which individuals differ (e.g., Swineford, 1938, 1941; Ziller, 1957). But the actual amount of guessing which an examinee does on a particular form of a test depends also on the form's explicit content (as compared with other equivalent forms of the same test), and on the specific time and circumstances under which it is administered. The guessing-tendency behavior has limited reliability, and varies about the true trait score with changes in test form and in occasions on which the test is given.

The content error e varies from form to form and from time to time, and so also does the guessing error. Let x_1 and x_2 be the raw scores on two forms of a test, t the content true score, g the guessing-tendency true score, e_1 and e_2 the content errors of measurement, and δ_1 and δ_2 the guessing-tendency errors of measurement. In place of (1) we then have

$$x_1 = t + g + e_1 + \delta_1, \quad (2)$$

$$x_2 = t + g + e_2 + \delta_2.$$

If the two forms are administered simultaneously (as, e.g., the odd and even items of one test), the errors will be form-associated errors only. If they are administered at different times, each type of error will include an occasion-associated error as well as a form-associated error.

If the two forms are equivalent, it is assumed that t and g are the same in both forms, that t and g are uncorrelated with e_1 , δ_1 , e_2 , and δ_2 , that e_1 and δ_1 are uncorrelated with e_2 and δ_2 , and that the two forms are equally reliable and equally variable, so that $\sigma_1^2 = \sigma_2^2 = \sigma_x^2$. It is not assumed that either t and g , or e_1 and δ_1 , or e_2 and δ_2 are uncorrelated. We, therefore, simply associate the two true scores and the two error scores of each pair in equations (2), so that

$$x_1 = (t + g) + (e_1 + \delta_1), \quad (3)$$

$$x_2 = (t + g) + (e_2 + \delta_2).$$

By the variance-ratio definition of reliability and the assumption of equal variance, we then have from either of equations (3),

$$R = \frac{\sigma_{t+g}^2}{\sigma_x^2}. \quad (4)$$

The correlation of x_1 with x_2 is

$$r_{12} = \frac{\Sigma(t+g)^2 + \Sigma(t+g)(e_2 + \delta_2) + \Sigma(t+g)(e_1 + \delta_1) + \Sigma(e_1 + \delta_1)(e_2 + \delta_2)}{N\sigma_1\sigma_2}$$

Under the equivalence assumptions the last three terms in the numerator vanish, $\sigma_1\sigma_2 = \sigma_x^2$, and

$$r_{12} = \frac{\Sigma(t+g)^2}{N\sigma_x^2} = \frac{\sigma_{t+g}^2}{\sigma_x^2}, \quad (5)$$

which is the same as (4).

The correlation of x_1 with $(t + g)$ is

$$r_{1(t+g)} = \frac{\Sigma(t+g)^2 + \Sigma(t+g)(e_1 + \delta_1)}{N\sigma_1\sigma_{t+g}}$$

The second term in the numerator vanishes, $\sigma_1 = \sigma_x$, and

$$r_{1(t+g)} = \frac{\sigma_{t+g}^2}{\sigma_x\sigma_{t+g}} = \frac{\sigma_{t+g}}{\sigma_x} \quad (6)$$

which by (5) is $\sqrt{r_{12}}$. We would arrive at the same result for $r_{2(1-g)}$.

As compared with free-answer tests, these results differ only in that the true score is the true content score *plus* the true guessing-tendency score. If the correction for guessing is not used we are simply measuring, with some error, a composite of the content knowledge or ability and the guessing-tendency trait.

When the correction for guessing is used, most of the systematic error represented by g in (2) and (3) is removed, but the guessing errors, δ_1 and δ_2 , remain: they are intrinsic to measurement with multiple-choice tests.

Instructions to limit guessing are peculiarly insidious. Partial knowledge of an item is real and substantial, and its use in answering multiple-choice tests always involves guessing. If an examinee can eliminate one or two wrong alternatives, and guesses among the remainder, the odds are in his favor. If he has a hunch, he should play it: hunches are right with frequency greater than chance. Under guess-limiting instructions, examinees whose guessing tendencies are high ignore them and receive additional credit for partial knowledge, while examinees with low guessing tendencies heed them and receive no such credit. When the correction formula is used, partial knowledge is credited *on the average* (Cureton, 1966), but to permit this, examinees must be instructed emphatically to omit an item *only* if an answer would be a *pure* guess.

If every examinee is *required* to mark every item, the g -variance is reduced to zero at the expense of an increase in the δ -variance.

REFERENCES

- Cureton, E. E. The correction for guessing. *Journal of Experimental Education*, 1966, 34, 44-47.
- Frery, R. B. Reliability of multiple-choice test scores is not the proportion of variance which is true variance. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 359-365.
- Swineford, F. The measurement of a personality trait. *Journal of Educational Psychology*, 1938, 29, 295-300.
- Swineford, F. Analysis of a personality trait. *Journal of Educational Psychology*, 1941, 32, 438-444.
- Ziller, R. P. A measure of the gambling response-set in objective tests. *Psychometrika*, 1957, 22, 289-292.

THE PROBABILITY OF MISCLASSIFICATION OF STUDENTS ON MULTIPLE CHOICE EXAMINATIONS¹

WALTER H. CARTER, JR.²

Medical College of Virginia
Health Sciences Center
Virginia Commonwealth University

THERE is now extensive literature on methods of obtaining a student's true score on a multiple choice examination (Calandra, 1941; Chernoff, 1962; Hamilton, 1950; Lord, 1953; Lysterly, 1951). Since, in many respects, such an examination is equivalent to random sampling, a majority of these procedures estimate this quantity statistically. This estimate is then used to classify the students in groups, such as honors, pass, fail, or A, B, C, D, F. However, since this grouping is based on a particular value of the estimator, itself a random variable, any such classification of students will be subject to error. Very little research has been devoted to determining the probability of misclassifying students as a result of using the various methods of estimating true scores.

In this paper the probability of the misclassification of students is developed by a method which is independent of the particular procedure used to determine the aforementioned groups. To obtain this quantity it is necessary to assume the existence of a set of questions which will accurately measure a student's ability to answer questions for which he does not know the correct answer.

Krutchkoff (1967) has defined the separation level of grades, e.g.

¹ This work was supported by a National Institutes of Health Institutional Research Grant (5P07FR00016).

² I am indebted to Mrs. Lillian Kornhaber, Department of Biometry, Medical College of Virginia for the computational assistance she provided in the preparation of the tables which appear in this article.

A, B, C, D, F, as the probability that a student with a higher grade actually knew the answers to more questions than the student with a lower grade in an attempt to justify the use of multiple choice examinations. This would appear, at first glance, to be related to a probability of misclassification and indeed a simple function of it shall be used as an approximation. To arrive at an expression for the separation level of grades Krutchkoff has made two assumptions:

1. Partial knowledge plays no role in a student's guess at the answer to a question for which he does not know the correct response.
2. The class of students taking the examination is homogeneous.

That the first assumption is too restrictive can be seen from the following hypothetical example. Consider a student who does not know the correct answer to a given question which contains five possible answers. As a result of partial knowledge of the subject matter, he is able to eliminate two of the possible answers as incorrect. Hence, this student is now able to guess the correct answer with probability $\frac{1}{3}$ instead of $\frac{1}{5}$. For this reason, it will be assumed that partial knowledge plays an important role in a student's response to multiple choice questions.

In what follows, a probability of misclassification will be derived which is based on each student's partial knowledge of the subject matter. It should also be noted that this derivation does not require assumption two above. Krutchkoff's results will then be compared to those obtained by the methods developed in this paper. To facilitate this comparison we adopt Krutchkoff's notation. Let

N = the total number of questions, each with r possible answers;

X = the proportion of subject matter known;

W = the number of answers known;

Y = the number of correct guesses;

Z = the total number of correct answers;

p = the probability of correctly guessing the answer to a question.

The assumption is made here, as in Krutchkoff's paper, that for each student the proportion of the subject matter known is almost normally distributed, that is, the proportion of the subject matter known follows a normal distribution truncated at zero and one. We choose to work with this truncation by concentrating the lower tail probability at zero and the upper tail probability at one. As a

result of this assumption it can be seen that NX is almost normally distributed with parameters μ and σ^2 . It is assumed that the probability, p , of correctly guessing the answer to a question is a random variable conditioned upon the student's knowledge of the subject matter as measured by W , the number of answers known. It is further assumed that the value of p for a given W follows a Beta distribution with parameters α and β ; where $\alpha = h(W)$ and $\beta = k(W)$. Both $h(W)$ and $k(W)$ denote unknown positive, real functions of W . It will not be necessary to estimate the functional form of these two functions, but for each student we shall arrive at an estimate of the value of α and β . The estimation of these parameters will be discussed later.

Theoretical Development

The conditional probability mass function of Z for given W and p can be expressed as

$$P(Z | W, p) = \binom{N-W}{Z-W} p^{z-w} (1-p)^{N-z}, \quad (1)$$

and from the definition of conditional probabilities we can write

$$P(Z | W) = \int_p P(Z | W, p) dF(p | W).$$

However, it has been assumed that

$$\frac{dF(p | W)}{dp} = \frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1}, \quad (2)$$

where

$$B(\alpha, \beta) = \int_0^1 p^{\alpha-1} (1-p)^{\beta-1} dp.$$

Hence,

$$P(Z | W) = \binom{N-W}{Z-W} \frac{B(Z-W+\alpha, N-Z+\beta)}{B(\alpha, \beta)}. \quad (3)$$

From this distribution it is possible to obtain an expression for the expected value of the total number of correct answers, $E(Z)$, as follows:

$$\begin{aligned}
 E(Z - W) &= E[E[(Z - W) | W]] \\
 &= E \left[\sum_{x=0}^{N-W} \int_0^1 \left\{ (Z - W) \binom{N-W}{Z-W} p^{x-W} (1-p)^{N-x} \right\} \right. \\
 &\quad \left. \cdot \frac{p^{x-1} (1-p)^{N-x-1} dp}{B(\alpha, \beta)} \right] \\
 &= E \left[\frac{(N - W) B(\alpha + 1, \beta)}{B(\alpha, \beta)} \right].
 \end{aligned}
 \tag{4}$$

If N is large and the truncated tails are small

$$EW \doteq \mu,$$

and

$$E(Z - W) = \frac{(N - \mu) B(\alpha + 1, \beta)}{B(\alpha, \beta)} \tag{5}$$

Together, the last two expressions imply

$$E(Z) = \frac{(N - \mu) B(\alpha + 1, \beta)}{B(\alpha, \beta)} + \mu, \tag{6}$$

from which it can be seen that

$$\mu = \frac{E(Z) - \frac{NB(\alpha + 1, \beta)}{B(\alpha, \beta)}}{1 - \frac{B(\alpha + 1, \beta)}{B(\alpha, \beta)}} \tag{7}$$

As a result of the distributional assumption made for X , it is easily seen that

$$P(W) = \begin{cases} \Phi\left(\frac{1 - \mu}{\sigma}\right), & W = 0 \\ \Phi\left(\frac{W + 1 - \mu}{\sigma}\right) - \Phi\left(\frac{W - \mu}{\sigma}\right), & W = 1, 2, \dots, N - 1 \\ 1 - \Phi\left(\frac{N - \mu}{\sigma}\right), & W = N, \end{cases} \tag{8}$$

where $\Phi(\cdot)$ represents the value of the standard normal cumulative distribution evaluated at (\cdot) .

Making use of the definition of conditional probability once again, we have

$$P(Z) = \sum_W P(Z | W)P(W) \\ = \sum_{W=0}^N \binom{N-W}{Z-W} \frac{B(Z-W+\alpha, N-Z+\beta)P(W)}{B(\alpha, \beta)}. \quad (9)$$

An expression for $P(Z | W)$, the probability that a particular student gave the correct answer to Z questions when he knew the answer to W questions has been obtained in equation (3). A quantity which will be very useful to us in calculating the probability of misclassification is $P(W | Z)$, the probability that a student knew the answer to W questions when he gave the correct answer to Z questions. It turns out that $P(W | Z)$ can be obtained from $P(Z | W)$ and $P(W)$ by means of Bayes' inversion rule which yields

$$P(W | Z) = \frac{P(Z | W)P(W)}{\sum_W P(Z | W)P(W)} \\ = \frac{\binom{N-W}{Z-W} B(Z-W+\alpha, N-Z+\beta)P(W)}{\sum_{W=0}^N \binom{N-W}{Z-W} B(Z-W+\alpha, N-Z+\beta)P(W)}. \quad (10)$$

As a result of (10) the Probability of Misclassification, PMC, can now be obtained as

$$PMC = 1 - \sum_{i=0}^{Z_1} P(W = i | Z_2) \sum_{j=0}^{i-1} P(W = j | Z_1). \quad (11)$$

The PMC is nothing more than the probability that, in a comparison between two students, the student who gave the correct answer to fewer questions, Z_1 , actually knew the correct answers to as many or more questions than the student who gave the correct answer to Z_2 questions ($Z_2 > Z_1$), i.e.

$$PMC = P[W_1 \geq W_2 | Z_2 > Z_1].$$

Before the PMC can be calculated we must obtain an estimator for σ , a parameter in the distribution of W . Since $Z = W + Y$, $\text{Var } Z$, the variance of Z , can be expressed as

$$\text{Var } Z = \text{Var } (W + Y) \\ = \text{Var } W + \text{Var } Y + 2 \text{Cov } (W, Y).$$

From the assumed distributional form of W it is known that

$$\text{Var } W = \sigma^2.$$

In order to complete the expression for $\text{Var } Z$, it is necessary to write

$$\text{Var } Y = E[\text{Var}(Y | W)] + \text{Var}[E(Y | W)] \quad \text{and} \quad (12)$$

$$\text{Cov}(W, Y) = E[E((W - \mu)(Y - (N - \mu)p) | W)].$$

Since the number of correct guesses equals the difference between the total number of correct answers and the number of known answers, it follows that

$$\begin{aligned} \text{Var}[Y | W] &= \text{Var}[(Z - W) | W] \\ &= E[(Z - W)^2 | W] - E^2[(Z - W) | W]. \end{aligned}$$

Evaluating these expressions we find

$$\begin{aligned} E[(Z - W)^2 | W] &= \sum_{z=W-0}^{N-W} (Z - W)^2 \binom{N-W}{Z-W} \frac{B(Z - W + \alpha, N - Z + \beta)}{B(\alpha, \beta)} \\ &= \int_p \left\{ \sum_{z=W-0}^{N-W} \frac{(Z - W)^2}{B(\alpha, \beta)} \binom{N-W}{Z-W} p^{z-W} (1-p)^{N-z} \right\} \\ &\quad \cdot p^{\alpha-1} (1-p)^{\beta-1} dp. \end{aligned}$$

For the binomial distribution

$$E(Z - W)^2 = (N - W)p(1 - p) + (N - W)p$$

and hence

$$\begin{aligned} E[(Z - W)^2 | W] &= \frac{(N - W)B(\alpha + 1, \beta + 1)}{B(\alpha, \beta)} + \frac{(N - W)B(\alpha + 1, \beta)}{B(\alpha, \beta)}. \quad (13) \end{aligned}$$

It has been shown previously, equation 4, that

$$E[Y | W] = E[(Z - W) | W] = \frac{(N - W)B(\alpha + 1, \beta)}{B(\alpha, \beta)}.$$

Therefore

$$\begin{aligned} \text{Var}[Y | W] &= \text{Var}[(Z - W) | W] \\ &= \frac{(N - W)}{B(\alpha, \beta)} \left[B(\alpha + 1, \beta + 1) + B(\alpha + 1, \beta) \right. \\ &\quad \left. - \frac{(N - W)B^2(\alpha + 1, \beta)}{B(\alpha, \beta)} \right] \quad (14) \end{aligned}$$

and

$$E[\text{Var}(Y | W)] = \frac{(N - \mu)}{B(\alpha, \beta)} \left[B(\alpha + 1, \beta + 1) + B(\alpha + 1, \beta) - \frac{(N - \mu)B^2(\alpha + 1, \beta)}{B(\alpha, \beta)} \right] - \left[\frac{\sigma B(\alpha + 1, \beta)}{B(\alpha, \beta)} \right]^2. \quad (15)$$

To complete the expression for $\text{Var } Y$, we must calculate $\text{Var}[E(Y | W)]$. Using the expression given in equation (4) for $E[Y | W]$ it can be shown that

$$\begin{aligned} \text{Var}[E(Y | W)] &= \text{Var} \left[\frac{(N - W)B(\alpha + 1, \beta)}{B(\alpha, \beta)} \right] \\ &= \left[\frac{B(\alpha + 1, \beta)\sigma}{B(\alpha, \beta)} \right]^2. \end{aligned} \quad (16)$$

Combining equations (15) and (16) in the manner prescribed by equation (12) yields

$$\begin{aligned} \text{Var } Y &= \frac{(N - \mu)}{B(\alpha, \beta)} \\ &\cdot \left[B(\alpha + 1, \beta + 1) + B(\alpha + 1, \beta) - \frac{(N - \mu)B^2(\alpha + 1, \beta)}{B(\alpha, \beta)} \right]. \end{aligned} \quad (17)$$

We proceed next to find an expression for the covariance between the number of answers known and the number of correct guesses, $\text{Cov}(W, Y)$. It follows from the definition of the covariance between two random variables that

$$\begin{aligned} \text{Cov}(W, Y) &= E[E((W - \mu)(Y - (N - \mu)p) | W)] \\ &= E \sum_p (W - \mu)(Y - (N - \mu)p) \binom{N - W}{Z - W} \\ &\quad \cdot \frac{p^{Z - W + \alpha - 1}(1 - p)^{N - Z + \beta - 1} dp}{B(\alpha, \beta)} \\ &= \frac{B(\alpha + 1, \beta)}{B(\alpha, \beta)} E[(W - \mu)(\mu - W)] \\ &= -\frac{B(\alpha + 1, \beta)\sigma^2}{B(\alpha, \beta)}. \end{aligned} \quad (18)$$

Finally, we are able to write

$$\begin{aligned} \text{Var } Z = \sigma^2 + \frac{(N - \mu)}{B(\alpha, \beta)} \\ \cdot \left[B(\alpha + 1, \beta + 1) + B(\alpha + 1, \beta) - \frac{(N - \mu)B^2(\alpha + 1, \beta)}{B(\alpha, \beta)} \right] \\ - \frac{2B(\alpha + 1, \beta)\sigma^2}{B(\alpha, \beta)}, \end{aligned} \quad (19)$$

which implies

$$\sigma = \sqrt{\frac{\text{Var } Z - \frac{(N - \mu)}{B(\alpha, \beta)} \left[B(\alpha + 1, \beta + 1) + B(\alpha + 1, \beta) - \frac{(N - \mu)B^2(\alpha + 1, \beta)}{B(\alpha, \beta)} \right]}{1 - \frac{2B(\alpha + 1, \beta)}{B(\alpha, \beta)}}}. \quad (20)$$

Before a value for the PMC can be calculated μ and σ must be estimated. An estimate of μ can be obtained by replacing $E(Z)$ in equation (7) by \bar{Z} and σ can be estimated by replacing $\text{Var } Z$ in equation (20) by s_z^2 , where

$$\bar{Z} = \frac{\sum_{i=1}^k Z_i}{k}, \quad (21)$$

$$s_z^2 = \frac{\sum_{i=1}^k (Z_i - \bar{Z})^2}{k - 1}, \quad (22)$$

and k equals the number of examinations given to determine a student's grade. If $k = 1$, then the examination must be randomly divided into $k_1 \geq 2$ different sub-examinations in order to apply methods developed in this section. A similar procedure, called the split-half technique, is frequently employed when measuring the reliability of an examination. See, for example, Rosinski and Hamilton (1966).

Estimation of α and β

In arriving at the expression for the PMC nothing was said about how to obtain values for α and β , the parameters which appear in the distribution of a student's guessing ability. In this section we shall use a method due to Weiler (1965) to estimate these parameters.

Since partial knowledge has been shown to play an important role in guessing and since guessing only occurs on questions for which the answer is unknown, it seems appropriate to include on the examination several questions chosen so that the students would not be expected to know the answer, but chosen in such a manner as to allow a student's partial knowledge to help in arriving at the guessed answer and then to infer from a student's performance on this sample of questions the parameters of the underlying Beta distribution. As a result of the examination it can be determined that 100P per cent of the time a student had a value of p above p_1 and another 100P per cent of the time had a value below p_2 . Hence, the following system of equations

$$\begin{aligned}\frac{1}{B(\alpha, \beta)} \int_0^{p_1} p^{\alpha-1} (1-p)^{\beta-1} dp &= 1 - P \\ \frac{1}{B(\alpha, \beta)} \int_0^{p_2} p^{\alpha-1} (1-p)^{\beta-1} dp &= P.\end{aligned}\tag{23}$$

These two expressions can be solved simultaneously for the unknown parameters α and β by use of Pearson's *Tables of the Incomplete Beta Function* (1934).

A Single Examination

The following example will illustrate the application of the procedure described in this paper to a single examination. Since the extension to a series of examinations is immediate, it will not be discussed further in this paper. The examination being analyzed is a section of a larger examination which was given to 127 first year medical students at the Medical College of Virginia in September 1968. In order to obtain estimates of the parameters of the underlying Beta distribution, the examination was randomly divided into two parts such that on each part there was an approximately equal number of questions, five on the first part and four on the second part, designed to measure a student's partial knowledge. Based on the students' performances, it was decided to pass those who answered more than eight of twenty-one questions correctly. Since there were five students who answered eight questions correctly and three who answered nine questions correctly, it was of interest to obtain the PMC for these students. These misclassification probabilities and those calculated by Krutchkoff's method appear in

Table 1. Notice that Krutchkoff's method only permits the calculation of a misclassification probability of a grade, e.g. between those students who correctly answered eight questions and those who answered nine, as opposed to the students within a grade.

Since students will generally perform differently on the set or sets of questions designed to indicate their degree of partial knowledge, it is possible, by applying the methods developed here, to calculate the PMC for students who have correctly answered the same number of questions. Thus, we now have a method for ranking such students. The PMC has been calculated for the students who scored eight and nine on the test and the results appear in Tables 2 and 3 respectively.

Conclusion

Since students, for various reasons, do not possess the same levels of partial knowledge, they guess the correct answer to questions not completely known with different frequencies. In the past either no attempt has been made to account for this effect or it has been assumed that all students, when faced with a question they do not know, guess the correct answer with equal probability, $1/r$. In this paper it has been assumed that the probability with which a person guesses the correct answer to a multiple choice question is a random variable that follows a Beta distribution with unknown parameters. A method is given for estimating these parameters. A procedure, using the estimates of these parameters, is developed to obtain the probability of misclassifying students based on their

TABLE 1*

Probabilities of Misclassifying Students who Correctly Answered Eight and Nine Questions

	9(5, 4, 20.0, 75.0)	9(7, 2, 20.0, 75.0)	9(7, 2, 4.8, 47.2)
8(6, 2, 4.8, 47.2) ^b	0.839	0.889	0.332
8(3, 5, 20.0, 75.0)	0.410	0.478	0.047
8(4, 4, 20.0, 75.0)	0.448	0.516	0.063
8(5, 3, 20.0, 75.0)	0.410	0.478	0.047
8(7, 1, 20.0, 75.0)	0.477	0.545	0.077

Krutchkoff's separation level = $1 - \text{PMC}(8, 9) = 0.558$ $\text{PMC}(8, 9) = 0.442$

* The entries in Tables 1, 2, and 3 are the PMC's, i.e. in the i, j position we have tabulated the probability that student i actually knew as many or more correct answers than student j .

^b 8(6, 2, 4.8, 47.2) denotes a student who answered eight questions correctly, six on one-half of the test and two on the other half, with parameters $\alpha = 4.8$ and $\beta = 47.2$ in the assumed Beta distribution.

TABLE 2
Probabilities Used to Rank Students who Correctly Answered Eight Questions

	8(6, 2, 4.8, 47.2)	8(3, 5, 20.0, 75.0)	8(4, 4, 20.0, 75.0)	8(5, 3, 20.0, 75.0)	8(7, 1, 20.0, 75.0)
8(6, 2, 4.8, 47.2)					
8(3, 5, 20.0, 75.0)		0.935	0.915	0.935	0.897
8(4, 4, 20.0, 75.0)			0.561	0.606	0.529
8(5, 3, 20.0, 75.0)				0.638	0.563
					0.529

Note.—The assumptions made in the derivation of Krutchkoff's separation level will not permit the calculation of a separation level for students who answered correctly the same number of questions.

TABLE 3

The Probabilities Used to Rank Students who Correctly Answered Nine Questions

	9(5, 4, 20.0, 75.0)	9(7, 2, 20.0, 75.0)	9(7, 2, 4.8, 47.2)
9(5, 4, 20.0, 75.0)		0.636	0.123
9(7, 2, 20.0, 75.0)			0.082

individual performances on multiple choice examinations. By means of illustration, it is further shown that this procedure can be used to rank students who gave the correct answer to the same number of questions.

REFERENCES

- Calandra, A. Scoring formulas and probability considerations. *Psychometrika*, 1941, 6, 1-9.
- Chernoff, H. The scoring of multiple choice questionnaires. *Annals of Mathematical Statistics*, 1962, 33, 375-393.
- Hamilton, C. H. Bias and error in multiple choice tests. *Psychometrika*, 1950, 15, 151-168.
- Hubbard, J. P. and Clemans, W. V. *Multiple choice examinations in medicine*, Philadelphia: Lea and Febiger, 1961.
- Krutchkoff, R. G. The separation level of grades on a multiple choice examination. *The Journal of Experimental Education*, 1967, 36, 63-68.
- Lord, F. M. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 1953, 18, 57-76.
- Lyerly, S. B. A note on correcting for chance success in objective tests. *Psychometrika*, 1951, 16, 21-30.
- Pearson, K. *Tables of the incomplete beta function*. Cambridge, England: University Press, 1934.
- Rosinski, E. F. and Hamilton, D. L. Examination procedures as part of a new curriculum. *Journal of Medical Education*, 1966, 41, 135-142.
- Weiler, H. The use of incomplete beta functions for prior distributions in binomial sampling. *Technometrics*, 1965, 7, 335-347.

NONPARAMETRIC ITEM EVALUATION INDEX¹

STEPHEN H. IVENS

College Entrance Examination Board
Atlanta, Georgia

THE purpose of this paper is to develop a nonparametric index for evaluating the effectiveness of dichotomously scored items that takes into account *both* the difficulty level and the discrimination of the item. Although Davis (1951) did not recommend combining both item characteristics into a single index, Findley (1956) defended the procedure. Indices such as the biserial r , point-biserial r , and D (Findley, 1956) are dependent on item difficulty, while the rank-biserial r developed by Cureton (1956) and extended by Glass (1965, 1966) is independent of item difficulties. The criterion upon which this new index is based is that the best possible item will have a difficulty of .5 and have perfect discrimination. The .5 difficulty level was chosen because this value maximizes the number of discriminations an item can make, and a test composed of such items will have the greatest overall validity and reliability (See Gulliksen, 1945, 1950; Lord, 1952; Cronbach and Warrington, 1952). Admittedly, however, there are instances when this criterion would not be appropriate for item selection.

Consider item i as one of k items in a test administered to N individuals. If these N individuals are ranked, either in terms of their total score or an outside criterion (with tied ranks randomly broken), from low to high, then the response vector X_i for item i would be

¹ The original suggestion for this index came from T. S. Briley, Florida State University. The author is indebted to J. M. Laible, Eastern Illinois University, and to J. K. Brewer, Florida State University, for many helpful suggestions given on the original draft of this paper.

$$X_i' = [\alpha_1, \alpha_2, \dots, \alpha_N] \quad (1)$$

where $\alpha_j =$ 1 if the j th individual passed the item

$\alpha_j =$ 0 if the j th individual failed the item.

Since the individuals were ranked from low to high, a corresponding vector of ranks, denoted R_N , would be

$$R_N' = [1, 2, 3, \dots, N]. \quad (2)$$

The scalar product $X_i' \cdot R_N = Y_i$ is the rank sum of those individuals who passed item i , namely,

$$Y_i = X_i' \cdot R_N = \sum_{j=1}^N j\alpha_j. \quad (3)$$

Let us define n_0 and n_1 as the number of individuals who failed and passed item i respectively, such that

$$N = n_0 + n_1, \quad (4)$$

and \bar{R}_N as the median rank of vector R_N ,

$$\bar{R}_N = (N + 1)/2. \quad (5)$$

An index, W_i , can now be defined as

$$W_i = Y_i + n_0\bar{R}_N, \quad (6)$$

which will increase in value as the discrimination of the item increases and as the difficulty of the item approaches .5.

There are a total of 2^N equally likely response vectors, X_i , as n_1 takes on the values from 0 to N . Thus, the expected value of W_i , denoted \bar{W} , is derived by summing equation 6 over all possible response vectors and dividing by 2^N . Hence:

$$\bar{W} = \frac{\sum W_i}{2^N} = \frac{\sum \left(\sum_{j=1}^N j\alpha_j + n_0 \left(\frac{N+1}{2} \right) \right)}{2^N}. \quad (7)$$

Since

$$\begin{aligned} & \sum \left(\sum_{j=1}^N j\alpha_j + n_0 \left(\frac{N+1}{2} \right) \right) \\ &= \frac{N(N+1)}{2} \sum_{n_0=1}^N \binom{N-1}{n_0-1} + \frac{N+1}{2} \sum_{n_0=1}^N n_0 \binom{N}{n_0} \\ &= N(N+1)2^{N-1}, \end{aligned} \quad (8)$$

equation 7 reduces to

$$\bar{W} = \frac{N(N+1)}{2}. \quad (9)$$

The use of W_i as an item index is not meaningful, however, unless we know the maximum and minimum values that W_i can achieve. These values can be obtained from the following inequalities;

$$\frac{N(N+1)}{2} - \{N + (N-1) + \dots + [N - (n_0 - 1)]\} \leq \sum_{i=1}^N j\alpha_i,$$

and

$$\frac{N(N+1)}{2} - (1 + 2 + \dots + n_0) \geq \sum_{i=1}^N j\alpha_i$$

which hold for each value of n_0 since $\sum_{j=1}^N j\alpha_j = N(N+1)/2$ if $\alpha_j = 1$ for each j . Upon simplification the above inequalities become

$$\frac{(N - n_0)(N - n_0 + 1)}{2} \leq \sum_{i=1}^N j\alpha_i \leq \frac{N(N+1) - n_0(n_0 + 1)}{2}.$$

Substituting from equation 6 yields

$$\frac{N^2 + N - (Nn_0 - n_0^2)}{2} \leq W_i \leq \frac{N^2 + N + (Nn_0 - n_0^2)}{2}.$$

When N is even, the maximum value for the upper bound of W_i is

$$W_{\max} = \frac{5N^2 + 4N}{8}. \quad (10)$$

This maximum is achieved when $n_0 = N/2$ where the first $N/2$ α_j 's are zero. The minimum value for the lower bound is

$$W_{\min} = \frac{3N^2 + 4N}{8} \quad (11)$$

and is achieved when $n_0 = N/2$ where the last $N/2$ α_j 's are zero.

When N is odd, the maximum and minimum values for W_i are

$$W_{\max} = \frac{(N+1)(3N+1)}{8} \quad (12)$$

and

$$W_{\min} = \frac{(N+1)(5N-1)}{8}. \quad (13)$$

The maximum value occurs when $n_0 = (N+1)/2$ or $(N-1)/2$ where the first $(N+1)/2$ or $(N-1)/2$ α_j 's are zero. Analogously the minimum value occurs when $n_0 = (N+1)/2$ or $(N-1)/2$ where the last $(N+1)/2$ or $(N-1)/2$ α_j 's are zero. Thus, W_i is a maximum when discrimination is perfect and difficulty is .5.

W_i is still not satisfactory as an index of an item's effectiveness, however, because repeated administrations of a given item would not yield comparable W_i values unless N is constant over all administrations. This dependence on N can be eliminated if we subtract from W_i the average value \bar{W} and divide the difference by the maximum W_i value minus \bar{W} . This new value, denoted S_i , equals the "status" of the items or symbolically

$$S_i = \frac{W_i - \bar{W}}{W_{\max} - \bar{W}}. \quad (14)$$

For computational purposes, the above expression for S_i can be simplified by referring to equations 6, 9, 10, and 12. When N is even

$$S_i = \frac{8Y_i - 4n_i(N+1)}{N^2} \quad (15)$$

and when N is odd

$$S_i = \frac{8Y_i - 4n_i(N+1)}{N^2 - 1}. \quad (16)$$

The mean of the distribution of S_i values, for each N , can be found by summing equation 14 over all possible response configurations and dividing by 2^N to give

$$\bar{S} = \sum S_i / 2^N = \sum \left(\frac{W_i - \bar{W}}{W_{\max} - \bar{W}} \right) / 2^N.$$

Substituting equations 8 and 9 in the above expression and simplifying yields

$$\bar{S} = \left(\frac{1}{W_{\max} - \bar{W}} \left(N(N+1)2^{N-1} - 2^N \cdot \frac{N(N+1)}{2} \right) \right) / 2^N = 0.$$

Referring to the definition of Y_i , equation 15 can be written

$$S_i = \frac{4}{N^2} \left(2 \sum_{j=1}^N j\alpha_j - n_i(N+1) \right) \quad (17)$$

and equation 16 can be written

$$S_i = \frac{4}{N^2 - 1} \left(2 \sum_{j=1}^N j\alpha_j - n_i(N+1) \right). \quad (18)$$

The variance of the distribution of S_i values, for N even, is equal to

$$\sigma_{S_i}^2 = \frac{16}{N^4 2^N} \sum \left(2 \sum_{j=1}^N j\alpha_j - n_i(N+1) \right)^2 \quad (19)$$

and for N odd, is equal to

$$\sigma_{S_i}^2 = \frac{16}{(N^2 - 1)^2 2^N} \sum \left(2 \sum_{j=1}^N j\alpha_j - n_i(N+1) \right)^2. \quad (20)$$

In order to obtain usable formulas for the variance of the distribution of S_i values it is necessary to simplify the expression

$$\sum \left(2 \sum_{j=1}^N j\alpha_j - n_i(N+1) \right)^2, \quad (21)$$

which is common to both equations 19 and 20. It can be shown² that expression 21 is equal to

$$\frac{N(N+1)(N-1)}{3} \cdot 2^{N-2}. \quad (22)$$

Substituting this value in equations 19 and 20 yields

$$\sigma_{S_i}^2 = \frac{4(N^2 - 1)}{3N^3}, \quad \text{for } N \text{ even}; \quad (23)$$

and

$$\sigma_{S_i}^2 = \frac{4N}{3(N^2 - 1)}, \quad \text{for } N \text{ odd}. \quad (24)$$

We have shown that the distribution of S_i , for each N , has a known variance and is symmetrical about zero with maximum and minimum values of one and minus one respectively. For each n_0 and n_1 , S_i is a linear transformation of Wilcoxon's rank sum statistic

² The derivation of equation 22 from equation 21 has been deposited with the National Auxiliary Publications Service, c/o CCM Information Corporation, 909 Third Avenue, New York, N. Y. 10022. Copies of this material may be obtained by citing document #NAPS-01602 and remitting \$5.00 for photocopies and \$2.00 for microfiche.

T (Wilcoxon, 1945). Summing over n_0 and n_1 , we see that the S_i distribution is linearly related to the sum of the various T distributions, holding N constant. In that the asymptotic distribution of T is normal (Bradley, 1960, p. 136), the asymptotic distribution of S_i is also normal. Thus the probability of a given S_i value differing from zero can be estimated by calculating S_i/σ_{S_i} and comparing the value to tables of the normal distribution for larger N .

The difference between S_i and the rank-biserial index, rb , reported by Glass (1965, 1966) is due to the effect of item difficulty on S_i . The two indices will yield identical values for N even, whenever n_0 equals n_1 , and for N odd, whenever n_0 equals $n_1 \pm 1$. Two advantages of S_i over rb are: (1) S_i equals zero whenever n_0 equals zero or N , while rb is undefined in these two situations; and, (2) S_i attains its maximum value of one from only one possible response configuration (two if N is odd), while rb attains a value of one under $N - 1$ possible response configurations.

In summary, S_i is an easily computed nonparametric index that:

1. is dependent on item difficulty and discrimination;
2. has a known range and variance;
3. has a significance test for its difference from zero;
4. can be meaningfully compared across different administrations of the same items; and,
5. can be computed by using either the total score of the test in which the item is contained or an outside criterion.

REFERENCES

- Bradley, J. V. Distribution-free statistical tests. WADD Technical Report 60-661, Wright Air Development Division, Wright-Patterson Air Force Base, 1960.
- Cronbach, L. J. and Warrington, W. G. Efficiency of multiple-choice tests as a function of spread of item difficulties. *Psychometrika*, 1952, 17, 127-147.
- Cureton, E. E. Rank-biserial correlation. *Psychometrika*, 1956, 21, 287-290.
- Davis, F. B. Item selection techniques. In E. F. Lindquist (Ed.), *Educational measurement*. Washington, D.C.: American Council on Education, 1951, Pp. 266-328.
- Findley, W. G. A rationale for evaluation of item discrimination statistics. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1956, 16, 175-180.
- Glass, G. V. A ranking variable analogue of biserial correlation: Implications for short-cut item analysis. *Journal of Educational Measurement*, 1965, 2, 91-95.

- Glass, G. V. Note on rank biserial correlation. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1966, 26, 623-631.
- Gulliksen, H. The relation of item difficulty and interitem correlation to test variance and reliability. *Psychometrika*, 1945, 10, 79-91.
- Gulliksen, H. *Theory of mental tests*. New York: John Wiley & Sons, 1950.
- Lord, F. M. A theory of test scores. *Psychometric Monograph*, 1952, No. 7.
- Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1945, 1, 80-83.

BAYESIAN TECHNIQUES FOR TEST SELECTION

W. PAUL JONES

New Mexico State University, Alamogordo

F. L. NEWMAN

University of Miami, Coral Gables

IN the recent literature related to statistical inference, an increasing amount of attention has been given to the potential of the techniques included under the rubric of Bayesian statistics (Meyer, 1966). While the basic Bayesian theorem dates back to 1763, an interest in applications of the theorem has come into prominence only in the past few years. A perusal of this Bayesian literature indicates that while there can be little question as to the validity of the theorem, its applications have generated a great deal of controversy (Binder, 1964).

The purpose of this article is not to present a detailed analysis of the rationale behind Bayesian inferential techniques. Several references are available which provide such information (Meyer, 1966; Binder, 1964; Edwards, Lindman, and Savage, 1963), and the reader desiring to explore in depth is referred to these sources. Rather, this article will present an example of Bayesian techniques for a specific problem, that of test selection. Horst (1966) indicated a need for new applications of decision theory to the area of psychological measurement and prediction, and this article is addressed to this need. The format of the article will be to present a hypothetical example of the application of Bayesian procedures in test selection and to discuss some of the ramifications of the method as applied to the example. It has been said that modern emphasis in Bayesian inference began with the publication in 1959 of Schlaifer's *Probability and Statistics for Business Decisions*, and the pro-

cedures in the example were adapted from procedures suggested by Schlaifer in a later work also oriented to business applications (1961).

The two primary advantages of Bayesian techniques are: (a) prior information about the problem under investigation is utilized in the decision making process, and (b) a specific probability statement can be applied to the results of the calculations. One of the major criticisms leveled against Bayesian procedures is based on this first advantage. Some critics insist that the utilization of prior information somehow lessens the objectivity of the decision making process. However, Meyer (1964) has noted that all decision making procedures, including hypothesis testing, contain some subjective underpinnings and added that someone with no previous knowledge about a problem is hardly equipped to investigate the problem. In terms of the second advantage, while students are prone to interpret all confidence intervals as if they were probability statements regarding a point estimator, such interpretations have no basis in the rationale of classical procedures.

Hypothetical Example

The Problem

A counselor is faced with a decision regarding the use of a short aptitude test as an aid in placement of students in remedial, regular, or honors classes. The test will be used only if it has correlation equal to or greater than .50 with the criterion (academic grades). From past experience with aptitude tests in other scholastic circumstances, the counselor's data suggest that the correlation between test and performance will be .60, and furthermore, that the interval between .50 and .70 would approximately constitute a plus or minus one sigma interval for the coefficient. To make the assumption of a normal prior distribution more tenable, all coefficients are transformed into Fisher's z functions. Thus, the prior distribution has a mean of .69 ($r = .60$, $z = .69$) and a sigma of .16 ($[z_{.70} - z_{.50}]/2 = .16$). The criterion, the z transform of .50, is .55.

The counselor then takes a sample of 50 randomly selected students and finds a correlation between aptitude test and performance of .71 (expressed as a z transform) with a standard error of $\sqrt{1/(50 - 3)} = .15$. The question then is whether a classical inferential process

serves the counselor better than a Bayesian inferential process in aiding the counselor's decision as to whether he should use the aptitude test.

The Classical Approach

Here the counselor simply contrasts the two z values (.71 and .55) to test the null hypothesis that there is "no difference" with P (Type I error) .05:

$$\frac{z - z_0}{\sqrt{1/(N - 3)}} = \frac{.71 - .55}{.15} = 1.061$$

The conclusion would be: "Do not reject the null hypothesis." Further, by computing a confidence interval and transforming the z functions back to correlation coefficients, the counselor may expand his nonrejection statement to say that the interval .40 to .76 has a .95 probability of covering the true predictive value. This is, however, the limit of inference allowed with this problem.

The Bayesian Approach

There are a number of inferential developments possible under a heading Bayesian inference. The one that seems most appropriate for this problem would be to first establish a posterior distribution based on the prior distribution and the sample, and to contrast it with the criterion of .55. From this procedure, a definite probability statement can be made regarding the criterion and the efficiency of the aptitude test.

In accordance with the formulation suggested by Schlaifer (1961), the mean (M_2) and standard deviation (sd_2) of the posterior distribution are:

$$M_2 = \frac{M_1(h_1) + M_s(h_s)}{h_1 + h_s} \quad (sd_2)^2 = \frac{1}{h_1 + h_s}$$

where

M_1 = prior mean (.69)

M_s = sample mean (.71)

h_1 = reciprocal of prior variance ($1/.16^2$)

h_s = reciprocal of sample variance ($1/.15^2$)

In this example the values are: M_2 of 0.70 and sd_2 of 0.11. The

posterior distribution and criterion can then be used to establish an explicit probability statement such as:

$$P\left[z < \left(\frac{.70 - .55}{.11} = 1.36\right)\right] = 0.087$$

The verbage of inference here takes a different tack than the classical procedure. Given the posterior distribution, the probability of obtaining a value z less than .55 is .087 or less than one in 10. Thus, the chances are better than nine out of 10 that the new test will have a predictive value equal to or greater than the criterion. The criterion value, .55, becomes the lower limit of prediction for the counselor's probability statement, rather than the pivotal quantity of an interval. The counselor is not accepting or rejecting the sample data, but rather incorporating the data in his probability statement. Now, if he chooses to use the aptitude test, the decision to do so would be justified by the probability statement, and not simply because the test is "no different" from the criterion.

According to a Bayesian view of the world, the sample data should now be incorporated into the counselor's information for further decision making. For example, if the counselor wishes to draw another sample of 50 and run the correlation again, he can use information from the first sample in evaluating results of future samples. Specifically, he can determine the critical value (C_v) below which the aptitude test's correlation with performance would be suspect. Before collecting data on a second sample, the C_v could be estimated (Schlaifer, 1961) as follows:

$$C_v = \frac{\text{criterion } (h_2 + h_s) - M_2(h_2)}{h_s}$$

$$= \frac{.55\left(\frac{1}{.11^2} + \frac{1}{.15^2}\right) - .70\left(\frac{1}{.11^2}\right)}{\frac{1}{.15^2}} = .27$$

Thus, given the posterior distribution (based on prior distribution and first sample), the counselor should not reject the aptitude test unless the second sample correlation is less than a z of .27.

Discussion

In order to use the procedures above, the only information neces-

sary that would not typically be used is the prior distribution of the variable. The other data will be available to the counselor, but he certainly should also have some previous data or prior belief about the relationship between aptitude tests and performance criteria. Only in Bayesian procedures can this previous knowledge be utilized in the statistical analysis.

The decision to use Fisher's z functions, while necessitating some adjustments in the prior distribution was necessary to justify the assumption of a normal prior distribution. While Bayesian techniques are certainly not limited to normal prior distributions, the calculations involved are simplified when a normal distribution can be assumed. Schlaifer (1961) has indicated that if the variance of the prior distribution is large compared with the variance of the sample, the mean and variance of the prior distribution can be substituted into the formulae which apply to normal prior distributions with no material loss in accuracy. This is evident in a perusal of the formula for determining the posterior mean which indicates that the posterior mean is a weighted average of the prior mean and the sample mean, and that the mean with the smallest variance receives the largest weight.

Therefore, with a minimal amount of computation, the counselor is able to synthesize his previous knowledge of the problem with data from the sample and determine a specific probability to attach to his decision. This alone would seem to justify the procedures, but as noted in the example, he can further use both his previous knowledge and results from the first sample in decisions regarding the results of still other samples.

Other applications of these procedures could include comparisons of different test instruments, an analysis of the value of adding more predictors, and various other problems of importance to the psychometrician. It is not suggested that Bayesian techniques will or should replace other inferential procedures. However, the techniques discussed in this article should be a useful addition to the statistical repertoire of personnel given responsibility for decisions regarding the use of various test instruments.

REFERENCES

- Binder, A. Statistical theory. *Annual Review of Psychology*, 1964, 15, 277-310.
Edwards, W., Lindman, H. and Savage, L. J. Bayesian statistical

inference for psychological research. *Psychological Review*, 1963, 70, 193-242.

Horst, P. *Psychological measurement and prediction*. Belmont, California: Wadsworth Publishing Co., 1966.

Meyer, D. P. A Bayesian school superintendent. *American Educational Research Journal*, 1964, 1, 219-222.

Meyer, D. P. Bayesian statistics. *Review of Educational Research*, 1966, 36, 503-515.

Schlaifer, R. *Probability and statistics for business decisions*. New York: McGraw-Hill, 1959.

Schlaifer, R. *Introduction to statistics for business decisions*. New York: McGraw-Hill, 1961.

PROBLEMS WITH INFERRING TREATMENT EFFECTS FROM REPEATED MEASURES¹

CHARLES E. WERTS AND ROBERT L. LINN

Educational Testing Service

A commonly encountered research design is that in which treatments are randomly assigned to available preformed groups and the differential effectiveness of the treatments is evaluated from measurements on the same instrument at the beginning and at the end of the experiment, it being known that the effect is appropriately indicated by changes in the group means. True experimental design in such a situation would require that more than one group be assigned at random to each treatment. The group could then be used as the unit of analysis and the differential treatment effects could be tested using the variation among group means within treatments.

In practice, however, there is frequently only one group per treatment and the individuals are used as the unit of analysis. Since the treatments are randomly assigned, either the analysis of covariance using initial status as the covariate (this would be "usage 2" mentioned by Evans and Anastasio, 1968) or a two factor (groups and time) analysis of variance for repeated measures (Winer, 1962) would appear relevant to the problem. In this paper the differences in interpretation that arise from using these two quasi-experimental procedures on the same data will be examined. While neither of these methods can be considered proper experimental design a comparison of the logic of the two procedures

¹ The research reported herein was performed pursuant to Grant No. OEG-1-6-061830-0650 Project No. 6-1830 with the Office of Education, U. S. Department of Health, Education, and Welfare. The assistance of Dr. Frederick Lord of Educational Testing Service is gratefully acknowledged.

provides a better understanding of the logic of some quasi-experimental procedures.

Methodology

Because our purpose is primarily to understand the logic behind the two methods, it shall be assumed hereafter that all measures are infallible and that random sampling procedures are perfect so that the treatment-covariate correlation is zero. It is also assumed that any extraneous variables that influence growth are unrelated to the influences being studied and that all relationships are linear additive.

A. The Analysis of Covariance (ANCOVA)

The mathematical model for ANCOVA is:

$$Y_{ij} = A_j + B_w X_{ij} + e_{ij} \quad (1)$$

where

Y_{ij} = final status,

X_{ij} = initial status,

B_w = pooled within groups regression slope,

A_j = intercept of the Y on X regression line for group j , i.e.,
 $A_j = \bar{Y}_j - B_w \bar{X}_j$ where \bar{Y}_j and \bar{X}_j are the respective Y
 and X means for group j , and

e_{ij} = error term assumed independent of A_j and X_{ij} and with zero mean.

ANCOVA requires homogeneity of within groups regression slopes, which if not found would mean that the treatments could not be simply ordered on a scale of effectiveness. Given this assumption, the special ANCOVA case in which the final status Y_{ij} is the initial status X_{ij} plus a growth component G_{ij} (\bar{G}_j is the mean growth for group j) may be delineated. It follows from equation (1) that $X_{ij} + G_{ij} = \bar{X}_j + \bar{G}_j - B_w \bar{X}_j + B_w X_{ij} + e_{ij}$, since $A_j = \bar{Y}_j - B_w \bar{X}_j$ and $\bar{Y}_j = \bar{X}_j + \bar{G}_j$.

Solution of the normal equations for B_w yields $B_w = 1 + B_{GX.A}$, where $B_{GX.A}$ is the within groups regression weight of growth on initial status and the one arises from the fact that part of final status is initial status. In this model, a nontrivial $B_{GX.A}$ indicates that the rate of growth is influenced by the level of the initial status. Substitution of B_w into the equation for the intercepts yields $A_j =$

$\hat{Y}_i - B_{wX} \hat{X}_i = \hat{X}_i + \hat{G}_i - (1 + B_{GX.A}) \hat{X}_i = \hat{G}_i - B_{GX.A} \hat{X}_i$. In other words, the intercept represents the mean growth for a group (\hat{G}_i) corrected for the net influence of X on growth, i.e., A_i is a measure of the net influence of treatments on growth with X_{ii} controlled.

In summary, when initial status is the covariate in the analysis of covariance, the "treatment effects" represent the net influence of groups on growth with initial status controlled. To the degree that the within groups slope differs from unity it may be inferred in this model that initial status also influences the rate of growth for individuals.

B. The Repeated Measures Analysis of Variance (Winer, 1962)

The analysis of variance (ANOVA) design relevant to our example can be represented schematically as:

	Time	
	Initial	Final
Treatments $j = 1$	Group 1	Group 1
$j = 2$	Group 2	Group 2
$j = 3$	Group 3	Group 3

In this analysis the "main effects" of treatments are completely confounded with differences between groups, however, the "main effects" of time as well as the treatment by time interaction are free of such confounding. Winer (1962) states: "The primary purpose of repeated measures on the same elements is the control that this kind of design provides over individual differences between experimental units. In the area of the behavioral sciences, differences between such units often are quite large relative to differences in treatments which the experimenter is trying to evaluate." Depending on the particular effect being studied the experimenter may or may not be interested in the time main effect, i.e., whether across all treatments there is on the average a net change in status, e.g., if the effect of diet were being studied on groups of children, the average gain in weight over time would be confounded with maturational trends and therefore of little interest. The *differential* impact of various treatments is indicated by the treatment by time interaction which is a measure of whether the mean changes differ among the treatments. Differential treatment effects in ANCOVA corre-

spend to variation among the intercepts, i.e., the A_i in equation (1).

Comparison of the Methods

A case in which three treatments have the same mean gain is depicted in Figure 1.

The repeated measures ANOVA would indicate no treatment by time interaction since the difference between initial and final status is the same for all groups. The condition under which ANCOVA would also indicate no differential treatment effects can be derived as follows:

$$\text{Since } A_i = \bar{Y}_i - B_w \bar{X}_i$$

$$\sigma_A^2 = \sigma_Y^2 + B_w^2 \sigma_X^2 - 2 B_w \sigma_{XY},$$

where σ_A^2 , σ_Y^2 , and σ_X^2 is the variance of the A_i , \bar{Y}_i , and \bar{X}_i , when these are assigned to individuals and

σ_{XY} = covariance of \bar{X}_i and \bar{Y}_i , when assigned to individuals. No differential effect implies that $\sigma_A^2 = 0$ and for the example in Figure 1 $\sigma_Y^2 = \sigma_X^2 = \sigma_{XY}$, therefore:

$$\sigma_Y^2 + B_w^2 \sigma_Y^2 - 2 B_w \sigma_Y^2 = 0,$$

or $B_w = 1$.

If $B_w = 1$ then $B_{GX.A} = 0$, i.e., initial status does not influence the rate of growth. In general, the difference between the ANOVA and

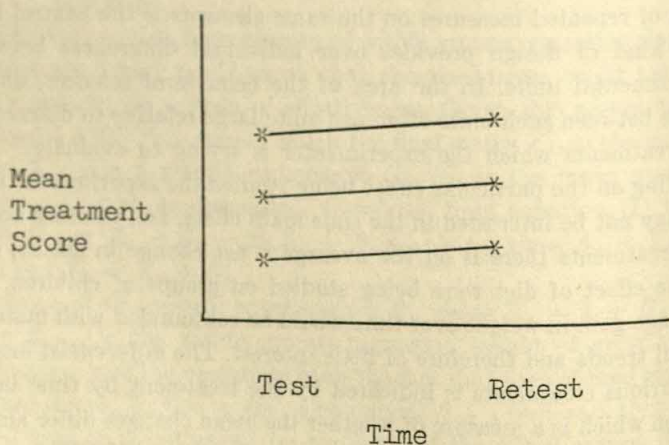


Figure 1. The case of equal gain.

ANCOVA models for simulating the action of a process is that the ANOVA assumes that initial status does not influence the rate of an individual's growth whereas the ANCOVA design allows this question to be settled by the within groups regression slope, i.e., by whether B_w differs from unity. This difference can be seen more clearly from Figure 2 which depicts the so-called "fanspread" hypothesis.

In this case the school means on retest correlate perfectly with the initial means but the former are more spread out. Because the differences between initial and final status are not the same for all groups the repeated measures ANOVA will indicate a treatment-by-time interaction. In contrast ANCOVA will not yield a differential treatment effect.

The within groups slope (which because of random assignment equals the between groups slope) for no differential treatment effect in ANCOVA can be derived as follows:

In the given example the correlation of \hat{X}_i and \hat{Y}_i is perfect, i.e.,

$$R_{\hat{X}\hat{Y}} = \frac{\sigma_{\hat{X}\hat{Y}}}{\sigma_{\hat{X}}\sigma_{\hat{Y}}} = 1$$

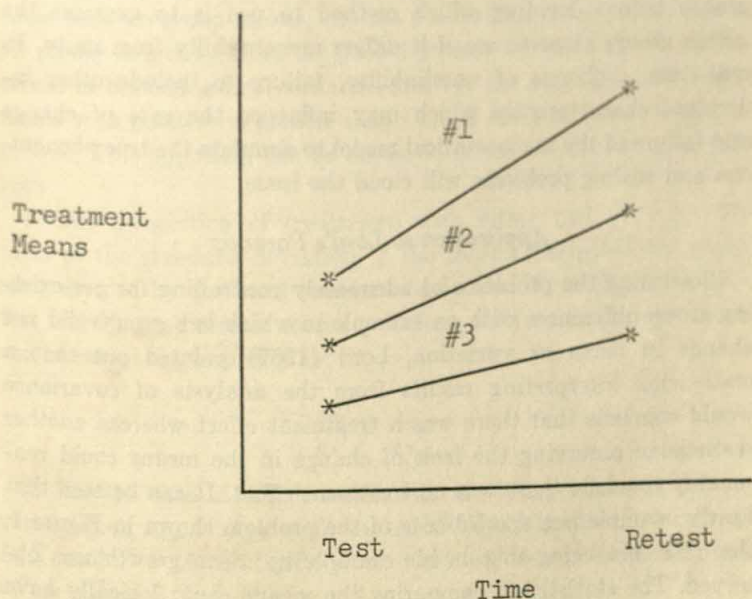


Figure 2. The Fanspread hypothesis.

Substituting into $\sigma_A^2 = \sigma_T^2 + B_w^2 \sigma_X^2 - 2B_w \sigma_{XT}$,
for $\sigma_A^2 = 0$ we obtain:

$$B_w = \frac{\sigma_T}{\sigma_X}.$$

In other words, if the within groups slope equals the (weighted) standard deviation of the final means divided by that of the initial group means then ANCOVA will indicate that the spreading apart of the school means is due to the influence of initial status on rate of gain rather than to a differential school effect. These results follow from the specification that the correlation between \bar{X}_i and \bar{Y}_i is perfect. If the correlation between \bar{X}_i and \bar{Y}_i is less than one then σ_A^2 will always be greater than zero.

Our analysis indicates that in a linear model the repeated measures ANOVA is an appropriate design for investigating differential treatment effects if initial status does not influence the rate of gain whereas ANCOVA would be appropriate if it does. The real problem is that in many studies the experimenter does not know apriori which is really the case. If the truth is unknown, then there is no way of knowing which method to believe. One possible approach before deciding which method to use is to examine the within groups slope to see if it differs meaningfully from unity. In real data, problems of unreliability, failure to include other individual characteristics which may influence the rate of change and failure of the mathematical model to simulate the true phenomena and scaling problems will cloud the issue.

Application to Lord's Paradox

Illustrating the problems of adequately controlling for pre-existing group differences with an example in which two groups did not change in mean or variation, Lord (1967) pointed out that a statistician interpreting results from the analysis of covariance would conclude that there was a treatment effect whereas another statistician observing the lack of change in the means could reasonably conclude there was no treatment effect. It can be seen that Lord's example is a special case of the problem shown in Figure 1, the difference being that in his example no mean growth was observed. The statistician comparing the means could logically have applied the repeated measures ANOVA. It follows that Lord's

statisticians implicitly had differing assumptions about the nature of reality which led to their contradictory interpretations. More precisely the ANCOVA user was in essence asserting that initial status does influence the rate of change in weight and that the deviation of the within groups slope from unity was an appropriate measure of this influence. In contrast the statistician who compared means was essentially asserting that initial status does not influence change irrespective of the fact that the within group slope was less than unity. These differing theories led to the opposite conclusions and there is no way of judging which was right since the truth is unknown. Too often researchers have forgotten that the inferences drawn from statistical analyses can be no more valid than the degree to which the mathematical model simulates reality. Missing from many studies is an examination of the model underlying the statistical measures in light of what is known or postulated about the phenomena under study.

Application to Campbell's Quasi-Experimental Approach

Given pretest and posttest data, Campbell and Clayton (1961) note several symptoms of treatment effects: (a) increases in differences between group means, (b) the posttest variance pooled across all groups is greater than the pooled pretest variance due to an increase in between school variance, and (c) the association of treatments with posttest is greater than that of the treatments with the pretest. This approach can be stated in terms of ANCOVA as follows:

1. The association of treatments with either test is "eta," the ratio of the standard deviation of the means (weighted by group size) divided by the total standard deviation. When X_{ij} is the pretest and Y_{ij} the posttest scores as before, then condition (c) above can be stated algebraically as:

$$\frac{\sigma_{\bar{X}}^2}{\sigma_X^2} < \frac{\sigma_{\bar{Y}}^2}{\sigma_Y^2}.$$

2. For the case in which treatments are independent of the covariate (i.e., $\sigma_{AX} = \sigma_{AX} = 0$) the ANCOVA formulas $\bar{Y}_i = A_i + B_w \bar{X}_i$ and $Y = A_i + B_w X_{ij} + e_{ij}$ yield: $\sigma_Y^2 = \sigma_A^2 + B_w^2 \sigma_{\bar{X}}^2$, and $\sigma_Y^2 = \sigma_A^2 + B_w^2 \sigma_X^2 + \sigma_e^2$.

3. By substitution

$$\frac{\sigma_T^2}{\sigma_X^2} < \frac{\sigma_A^2 + B_w^2 \sigma_X^2}{\sigma_A^2 + B_w^2 \sigma_X^2 + \sigma_e^2}.$$

4. Examination of this formula shows that if there is no differential treatment effect (i.e., $\sigma_A^2 = 0$) then $(B_w^2 \sigma_X^2) \div (B_w^2 \sigma_X^2 + \sigma_e^2)$ will always be smaller than $\sigma_X^2 \div \sigma_X^2$. When the inequality is satisfied then $\sigma_A^2 > 0$. From the perspective of ANCOVA this criterion is quite stringent since it can easily be shown that σ_A^2 can be nonzero when the inequality is not satisfied. When σ_e^2 is large then σ_A^2 must be correspondingly large to be detected using the inequality criterion.

5. When Campbell's paradigm is applied to cases where treatments are not assigned randomly as in naturalistic school effects studies, then the formulas become more complicated since $\sigma_T^2 = \sigma_A^2 + B_w^2 \sigma_X^2 + 2B_w \sigma_{AX}$ and $\sigma_Y^2 = \sigma_A^2 + B_w^2 \sigma_X^2 + 2B_w \sigma_{AX} + \sigma_e^2$ where $\sigma_{AX} = \sigma_{AX}$. It is still true that $\sigma_A^2 > 0$ when the inequality is satisfied but $\sigma_A^2 > 0$ may be true even when the inequality is not satisfied especially in some of the cases when B_w or σ_{AX} are negative.

6. The researcher who uses the Campbell-Clayton approach has, like Lord's statistician who used ANCOVA, assumed that initial status can influence growth. Furthermore, this approach is unsatisfactory for detecting influences which tend to reduce differences between groups such as compensatory education efforts or social pressures towards societal norms.

The example used in this paper in which treatments were assigned randomly to preformed groups can be classified as a multi-group variant of Campbell and Stanley's (1963) design No. 10, the nonequivalent control group design. A control group receiving no treatment may for statistical purposes be handled as another treatment group. The notion that initial status may influence growth is an instance of what Campbell and Stanley call the "interaction of selection and maturation." ANCOVA in essence tries to rule out one type of selection-maturation interaction by using the within groups slope to estimate the effect of differential selection (i.e., mean initial status) on maturation. It is known that ANCOVA results become difficult to interpret when within group homogeneity of regression is not found, (which might mean that whether one treatment is better than another depends on which subject it is applied to). In these cases the notion of a treatment effect applicable to everybody in a group is no longer appropriate, yet such a notion

is probably implicit in most uses of ANCOVA and the Campbell-Stanley design #10. When homogeneity of within group regression is found ANCOVA rules out the selection-maturation hypothesis that the change in means is proportional to initial status. If the change in means is proportional to initial status the increase in the within group variance will be in general proportionately larger than the increase in the between groups variance since σ_e^2 is greater than zero. In such cases the treatment-test correlation will decrease from test to retest, i.e., neither the Campbell-Clayton criteria nor ANCOVA will yield a differential treatment effect.

Comparison of the ANCOVA to the ANOVA model

A major difference between the ANCOVA and the ANOVA designs is that the latter does not interpret the within groups covariance between initial and final status. Therefore, it is consistent with the logic of ANOVA to measure only a random sample of the members from each group on pretest and a new random sample from that group on posttest. Similarly it would be reasonable in ANOVA to randomly split the members of each group such that half received the pretest and half the posttest, as a means of avoiding practice effects. Thus Schaie's (1965) model for studying developmental effects shares the logic of the ANOVA model with respect to the influence of initial status.

The neglect of the within groups covariance suggests that from the viewpoint of the general linear model the test-retest ANOVA design may be considered as a special case of the ANCOVA model. Consider the dummy variable form of ANCOVA:

$$Y_{ij} = B_0Z_0 + B_1Z_1 + \cdots + B_{J-1}Z_{J-1} + B_wX_{ij} + e_{ij}$$

where

Z_0 = dummy code of 1 for everybody, and

Z_j = 1 for persons in group j , 0 for others

(J = total number of groups).

If B_w were unity then the X_{ij} term could be shifted to the left side of the equation yielding:

$$Y_{ij} - X_{ij} = G_{ij} = B_0Z_0 + B_1Z_1 + \cdots + B_{J-1}Z_{J-1} + e_{ij}$$

This equation is simply the analysis of variance of change scores which yields the identical differential treatment effect indicated

by the treatments by time interaction in the repeated measures design. Whereas the ANCOVA model is the special case of the linear model with homogeneous within group regression, the treatment by time interaction of ANOVA analysis of test-retest measures may be considered the special case of the repeated measures ANCOVA model with the additional requirement that initial status not influence growth, i.e., $B_w = 1$.

REFERENCES

- Campbell, D. T. and Clayton, K. N. Avoiding regression effects in panel studies of communication impact. *Studies in Public Communication*, University of Chicago Press, 1961, No. 3, 99-118.
- Campbell, D. T. and Stanley, J. S. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963, pp. 171-246.
- Evans, S. H. and Anastasio, E. J. Misuse of analysis of covariance when treatment effect and covariate are confounded. *Psychological Bulletin*, 1968, 69, 225-234.
- Lord, F. M. A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 1967, 68, 304-305.
- Schaie, K. W. A general model for the study of developmental problems. *Psychological Bulletin*, 1965, 64, 92-107.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962, Chapter 7.

ARE THERE TWO EXTREMENESS RESPONSE SETS?

LEONARD V. GORDON

State University of New York at Albany

THE extremeness response set is defined as the individual's tendency to use the more extreme scale alternatives when responding along an intensity dimension, such as one ranging from "Strongly Agree" to "Strongly Disagree." Since this response set has been assumed to be independent of the scale content itself, some investigators have recommended its elimination through the use of an appropriate scoring scheme, while others have taken it to be a worthwhile subject of study in its own right.

Hamilton (1968) in a comprehensive review of the literature relevant to the extremeness response set concluded that it can be reliably measured but described the evidence regarding its correlates as contradictory or at least unclear. He ascribed this lack of agreement largely to the diversity of scale content and the different techniques for extremeness response set measurement employed in the studies reviewed. Another possibility, apparently not considered by Hamilton, is that there are two extremeness response sets, one at each pole, which are sufficiently different from one another to require separate summarizations. His own observations that the correlation between the measures at the two poles is substantially lower than the response set reliabilities suggest that this may very well be the case.

That the tendency to mark extreme responses at the positive and negative poles of a response continuum have substantially different characteristics was incidentally observed in three studies by the writer: (a) In the first (Gordon, 1952) three forms of a 150-item personality inventory, employing five-choice response format, were randomly administered to large samples of subjects, with the same

30-items appearing in the first, middle and last sections on different forms. It was consistently found that the later the same socially undesirable items appeared in the questionnaire the greater the number of extreme negative endorsements they received. On the other hand, no counterpart increase in extreme positive endorsements of socially desirable items was noted. (b) In the second (Gordon, 1967a) the validities of various scoring schemes for the Work Environment Preference Schedule (Gordon, 1968a), a 5-point Likert scale designed to measure the "bureaucratic personality," were compared. Responses at the disagreement end of the continuum were found to contribute differentially to scale variance, while those at the agreement end did not. The scoring scheme which differentially weighted extremeness responses only at the negative pole yielded the highest validities against relevant external criteria. (c) In the third study (Gordon and Kikuchi, 1970), a school form of the Work Environment Preference Schedule in the original and in translation was administered respectively to counterpart American and Japanese high school students. In both cultures and for both sexes, significant differences were found between the contributions of the response set scores at the negative and positive poles to total scale variance, that for the negative pole being substantially larger.

The present study was conducted to further examine the validity of the extremeness response tendency at each of the two poles—employing two widely used research instruments, the California Authoritarianism or *F*-scale (Adorno, 1950) and the Dogmatism or *D*-scale (Rokeach, 1960). The design follows that of Korn and Giddan (1964) who used external measures of relevance to the construct validity of the Dogmatism scales as criteria. The expectation, based on prior findings, is that the two response set scores, one based on positive and one based on negative responses, will differentially contribute to scale validity and that the negative response set score will make the greater contribution.

Procedure

The California *F*-scale and Rokeach's *D*-scale each employ response alternatives representing three degrees of agreement and disagreement: "Very Strongly Agree" (VSA), "Strongly Agree" (SA), "Agree" (A), "Disagree" (D), "Strongly Disagree" (SD)

and "Very Strongly Disagree" (VSD). Six scores were obtained for purposes of the present analysis. The weights for each are presented in the rows of Table 1.

The *C* score is based on the conventional weighting scheme; the *P* score utilizes the dichotomous weights recommended by Cronbach (1950), Peabody (1962) and Korn and Giddan (1964) to eliminate extremeness response set variance; the *H* and *L* scores employ weights designed to capitalize on response set variance respectively at the "agree" and "disagree" ends of the continuum (Gordon, 1967a); the *X* and *Y* scores are response set scores which respectively measure the tendency to use the more extreme alternatives when agreeing or when disagreeing. Each response set score is computed by obtaining the weighted sum of the individual's responses at the appropriate half of the continuum and by then dividing by the number of such responses.¹

The external criterion scales employed in this study were conformity and Independence as measured by the Survey of Interpersonal Values or SIV (Gordon, 1960) and Variety and Orderliness as measured by the Survey of Personal Values or SPV (Gordon, 1967b). Selection of these scales was based on two considerations. First, the SIV and SPV utilize forced-choice format and thus are free of the particular response set under investigation. Second, and most important, these four scales as well as the Authoritarianism and Dogmatism scales had been found in a number of studies to be associated with bureaucratic personality structure (Gordon, 1968b, 1970). Thus, individuals who score high on Au-

TABLE 1

Weights Applied to Response Alternatives for the Six Scores

	VSA	SA	A	D	SD	VSD
C	3	2	1	-1	-2	-3
P	1	1	1	0	0	0
H	3	2	1	0	0	0
L	0	0	0	-1	-2	-3
X	3	2	1			
Y				1	2	3

¹ The *X* and *Y* scores differ computationally from the *H* and *L* scores in that for the latter the total number of responses (in this case 24) constitutes the divisor.

thoritarianism and Dogmatism would be expected to place a high value on conformist behavior and on being systematic and orderly, and a low value on independence of action and on openness to new and varied experiences, as assessed respectively by the four value scales.

The *F*-scale, *D*-scale, SIV and SPV were administered to a sample of 212 students primarily in the upper grades of a university high school, and product moment correlations among the several scores were obtained.

Results

A positive correlation between the *X* and *Y* scores serves as evidence for the existence of an extremeness response set since in unidirectional scales *more* extreme responses in one direction would be expected to be associated with *less* extreme responses in the other if content were the only consideration. In the present instance, the correlations between the *X* and *Y* scores were .48 and .63 for the Authoritarianism and Dogmatism scales respectively, reflecting a decided extremeness response set tendency on the part of these subjects.

Product moment correlations of the *X* and *Y* scores with the *C* and *P* scores for both the Authoritarianism and Dogmatism scales are presented in Table 2. It will be noted that the *X* and *Y* scores are not independent of scale content, but in all instances are significantly related to the *C* and *P* scores of their respective instruments. That the relationships with the *P* score are significant is of interest since the latter score is designed to be extremeness-response-set free.² That the correlations with the *C* score are higher than

TABLE 2

Correlations of the X and Y Scores with the C and P Scores of the Authoritarianism and Dogmatism Scales

	Authoritarianism			Dogmatism	
	C	P		C	P
X	.26**	.14*	X	.21**	.16*
Y	-.30**	-.25**	Y	-.27**	-.14*

* $p < .05$.

** $p < .01$.

² The confounding effects of content and style noted by Hamilton (1968) is illustrated here.

those with the *P* score would be expected since the *C* score gives added weight to extreme responses. The relationships of the *Y* scores with the total scale scores are somewhat higher than those of the corresponding *X* scores, however none of the differences are statistically significant.

Product moment correlation of the several types of scale scores (*C*, *P*, *H*, and *L*) and two response set scores (*X* and *Y*) with the four forced-choice criterion scales are presented in Table 3. It will be noted that the four Authoritarianism scores are significantly related in a positive direction to Conformity and Orderliness and in a negative direction to Indifference and Variety. The correlations for the four Dogmatism scores are directionally identical, but of uniformly smaller magnitude. All relationships are in the hypothesized direction.

A comparison of the relationships of *X* and *Y* response set scores with each of the criterion scales reveals that in seven of the eight cases the absolute value of the correlation of the *Y* score is the higher. However, more important, in all eight instances the signs of the *Y* score correlations are directionally congruent with both theoretical expectation and the obtained *F* or *D* scale validities, while in six out of eight instances the signs of the *X* score correlations are in directions opposite to what would be anticipated on these bases. For example, individuals who are the more extreme when agreeing (*X*) with Authoritarianism items would be expected to score lower in Indifference—yet the obtained correlation is

TABLE 3

Correlations of the Six Authoritarianism and Dogmatism Scores with the Four Criterion Scales

Criterion	<i>X</i>	<i>Y</i>	Authoritarianism				<i>t_{H-L}</i>
			<i>C</i>	<i>P</i>	<i>H</i>	<i>L</i>	
Conformity	.02	-.13	.37**	.37**	.30**	.32**	.40
Indifference	.09	.19**	-.31**	-.30**	-.20**	-.36**	3.20**
Variety	.04	.09	-.35**	-.38**	-.24**	-.30**	1.17
Orderliness	.13	-.09	.37**	.40**	.32**	.33**	.20
Criterion	<i>X</i>	<i>Y</i>	Dogmatism				<i>t_{H-L}</i>
			<i>C</i>	<i>P</i>	<i>H</i>	<i>L</i>	
Conformity	-.11	-.18*	.10	.06	.01	.18*	2.81**
Indifference	.14*	.24**	-.19**	-.13	-.03	-.28**	4.27**
Variety	.14*	.16*	-.15*	-.16*	-.05	-.22**	2.83**
Orderliness	-.02	-.10	.15*	.14*	.11	.17*	.98

* $p < .05$.

** $p < .01$.

positive in sign (.09); those who are the more extreme when disagreeing with Authoritarianism items (those with higher Y scores) would be expected to score higher in Independence, and the relationship is positive (.19).

While the difference in directional congruence of the X and Y score validities is pronounced, an overall test of significance is precluded since the 16 response set coefficients are not independent. However, the differential influence of the X and Y response sets on scale validity may be assessed indirectly by comparing the validities of the H and L scores, since each of the latter scores is weighted so as to capitalize on response set variance at its respective end of the continuum. It will be noted (Table 3) that in the two instances where the validities of the X scores and the scale scores are congruent in sign (on Authoritarianism, with Conformity and Orderliness), validities of the H and L scores are not significantly different. Out of the six remaining instances where the signs of the validities of the X score and the scale scores are not congruent, in four cases the validities of the L scores are significantly higher ($p < .01$) than those of the corresponding H scores.³

Discussion

The results of the present study indicate that the scoring scheme that included only negative response set variance on the whole yielded significantly higher validities than that which included only positive response set variance. Thus, it may be inferred that the extremeness response sets at the two ends of the continuum differentially contribute to scale validity. One other study specifically designed to test this latter hypothesis was noted in the literature. Mitzel, Rabinowitz and Ostreicher (1956), using the Minnesota Teacher Attitude Inventory (MTAI) as their research instrument, supervisory ratings of teachers as the criterion, and response set measures statistically identical to the X and Y scores, observed that only the negative response set had significant validity. "The negative intensity response set was found to influence the test scores in such a way that test validity is increased by its presence. Positive intensity . . . was found to exert very little effect on MTAI validity" (p. 514).

That there are two functionally different extremeness response

³ Hotelling's test of significance for correlated correlations was employed (Guilford, 1965).

sets is supported by the weight of available evidence based on external validation, internal analysis, and correlations between the response set measures themselves. Thus, whether extremeness response sets are studied in their own right or to assess their effect on scale validity, separate analyses would certainly appear to be called for. Failure to do so may preclude the making of proper inferences regarding their characteristics or their influence.

REFERENCES

- Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., and Sanford, R. N. *The authoritarian personality*. New York; Harper, 1950.
- Cronbach, L. J. Further evidence on response sets and test design. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1950, 10, 3-31.
- Gordon, L. V. The effect of position on the preference value of personality items. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1952, 12, 669-676.
- Gordon, L. V. *Survey of interpersonal values*. Chicago: Science Research Associates, 1960.
- Gordon, L. V. Validity of scoring methods for bipolar scales. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 1099-1106. (a)
- Gordon, L. V. *Survey of personal values*. Chicago: Science Research Associates, 1967. (b)
- Gordon, L. V. *Work environment preference schedule*. Albany: Author, 1968. (a)
- Gordon, L. V. *Correlates of bureaucratic orientation*. In *Proceedings of XVIth International Congress of Applied Psychology*. Amsterdam: 1968, 291-297. (b)
- Gordon, L. V. Measurement of Bureaucratic Orientation. *Personnel Psychology*, 1970, 23, 1-11.
- Gordon, L. V. and Kikuchi, A. Response sets of Japanese and American students. *Journal of Social Psychology*, 1970, 82, 143-148.
- Guilford, J. P. *Fundamental statistics in psychology and education*. New York: McGraw Hill, 1965.
- Hamilton, D. L. Personality attributes associated with extreme response style. *Psychological Bulletin*, 1968, 69, 192-203.
- Korn, H. A. and Giddan, N. S. Scoring methods and construct validity of the dogmatism scale. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1964, 24, 867-874.
- Mitzel, H. E., Rabinowitz, W., and Ostreicher, L. M. The effects of response sets on the validity of the Minnesota Teacher Attitude Inventory. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1956, 16, 501-515.
- Peabody, D. Two components in bipolar scales. *Psychological Review*, 1962, 69, 65-73.
- Rokeach, M. *The open and closed mind*. New York: Basic Books, 1960.

PREDICTION OF INDIVIDUAL STABILITY¹

GEORGE V. C. PARKER

University of Texas at Austin

IN the history of the study of individual differences, one consistent observation has been that many organisms, human *Ss* in particular, are not consistent in their responses to repeated presentations of the same situations. The behavioral instability phenomenon has been rather problematic from both a theoretical and a practical standpoint. With his concept of behavioral oscillation, for example, Hull (1952) explicitly and formally recognized the theoretical significance of intraindividual variability but experimenters have typically averaged their measures, thus avoiding a direct confrontation with the issue of within-subject variability. In the area of test construction, response instability has traditionally been categorized as a kind of nuisance "error" to be minimized in order to achieve test reliability. Mischel's (1968) review of the assessment literature suggests that attempts to demonstrate reliability of "traits" via personality tests has not been entirely successful; he found substantial reliability only for tests of intellectual and cognitive style. Explanations for response inconsistency are many and have been with us for some time. For example, 35 years ago, Lentz (1934) attributed test response variability to such vicissitudinous characteristics as *S* indifference, as well as deficiencies in areas such as sincerity, sympathy, and appreciation of scientific method. While Lentz is doubtless not the first psy-

¹ This study was supported in part by a grant from the University Research Institute of The University of Texas at Austin. Protocol scoring and all computations were carried out with the CDC 6600 computer of The University of Texas at Austin. The author is grateful to Quinn McNemar for his generous advice concerning the data analyses, and to Gary Jacobsen and Eric Haufrecht for assistance with data analyses.

chologist who has charged *Ss* with not caring, it is equally certain that our explanatory wherewithal is not greatly improved today. For example, Block (1968) has recently examined some of the psychometric and psychological reasons for behavioral inconsistency. While rather gloomily noting the lack of inspiring empirical evidence for personality consistency, his speculations about the causes of inconsistency are not greatly different from Lentz's.

An alternative approach to the problem of intraindividual variability/stability is to consider it as a lawful phenomenon open to systematic investigation in its own right. Data from available studies, such as Fiske (1957), Berdie (1961), Baumeister and Kellas (1968), Carrigan (1963), and Worell (1963) suggest that this approach may be fruitful. Indeed, it seems likely that both theoretical and experimental understanding of the phenomenon of individual differences in stability could improve the validity of the prediction of the behavior of individuals. The present study, therefore, had the primary objective of evaluating further the usefulness of conceptualizing intraindividual stability as a personality construct. This was done by assessing the feasibility of developing a psychometric scale which might be used to predict reliably individual stability in self-description, self-concept, and other behavioral domains. Secondary objectives were to explore other concomitants of the scale.

Method

Development of the Stability scale (*Stb*) was based on the Gough Adjective Check List (ACL) which has shown considerable promise as a research instrument (Gough and Heilbrun, 1965; Parker, 1969).

Criterion Subjects and Procedure

One hundred forty male and 140 female undergraduates volunteering from introductory psychology classes were given the ACL under standard conditions a total of three times. Following the first ACL administration, the procedure was repeated twice at two-week intervals, providing three ACL protocols in a test-retest fashion for each *S*.

Intraindividual Stability Criterion Measure

After scoring the protocols for the 24 standard ACL scales (Gough and Heilbrun, 1965), the variance of each scale score was computed for each *S*. The average of these scale variance scores for each *S* furnished his overall index of self-descriptive stability. For the sexes separately, *Ss* were grouped into quartiles according to the distribution of stability indices, providing high- and low-stability criterion groups for ACL item-analysis of endorsement differences by *z* tests (McNemar, 1962, p. 60). The study of patterns of self-description associated with stability was undertaken by sexes separately because earlier work (Parker, 1969) has shown systematic sex-related differences in patterns of ACL item endorsement.

Results

For male *Ss*, Table 1 lists the adjectives for which stability criterion groups differed in rates of endorsement. These items com-

TABLE 1
*ACL Stability Scale Adjectives for Males**
Indicative Adjectives (37 Items)

Adjective	Proportion of High- Stability Males Endorsing Item	Proportion of Low- Stability Males Endorsing Item	Adjective	Proportion of High- Stability Males Endorsing Item	Proportion of Low- Stability Males Endorsing Item
alert	91	63	pleasant	86	60
attractive	60	34	poised	49	23
clear-thinking	94	66	practical	94	71
clever	74	49	progressive	83	40
confident	80	46	quick	71	31
conscientious	83	57	rational	83	57
determined	89	49	reasonable	100	71
dignified	63	34	relaxed	71	43
dominant	49	23	reliable	94	69
efficient	94	54	resourceful	77	51
enterprising	63	31	responsible	57	63
foresighted	69	37	sharp-witted	57	29
individualistic	83	57	sly	23	3
industrious	77	40	stable	80	46
initiative	80	26	tactful	80	49
opportunistic	57	31	thorough	57	26
organized	71	37	tolerant	74	49
original	63	31	versatile	80	51
planful	60	29			

* Differences significant $p < .05$.

TABLE 1 (Continued)
*ACL Stability Scale Adjectives for Males**
Contraindicative Adjectives (18 Items)

Adjective	Proportion of High-Stability Males Endorsing Item	Proportion of Low-Stability Males Endorsing Item
awkward	03	34
careless	09	43
coarse	03	19
complaining	11	40
confused	11	51
dreamy	23	51
emotional	40	69
forgetful	16	51
fussy	03	31
gloomy	06	29
impulsive	17	54
lazy	17	57
moody	40	74
nervous	23	54
shy	29	54
slow	00	37
sulky	06	26
worrying	20	57

* Differences significant $p < .05$.

prise the ACL *Stb* scale for males. Because there are 37 high-stability and 18 low-stability items, the ACL *Stb* scale scoring was organized in the indicative-minus-contraindicative manner typical of most other ACL scales. Thus, an individual's ACL *Stb* scale Raw Score = Σ High-Stability items minus Σ Low-Stability items. The average *Stb* raw score for an independent sample of 35 high-stability males = 19.7; for 35 low-stability males = 6.7; $F = 55.8$; $df = 1,68$, $p < .01$. This compares closely to the data from the original sample males: high-stability = 20.2, $N = 35$; low-stability = 6.9, $N = 35$; $F = 51.5$, $df = 1,68$, $p < .01$.

For female *Ss*, Table 2 lists the adjectives on which the stability criterion groups differed in rates of endorsement. These items comprise the ACL *Stb* scale for females, and would be scored in an indicative-minus-contraindicative fashion as described above for males. The average *Stb* raw score for an independent sample of 36 high-stability females = 5.3; for 38 low-stability females = 9.4; $F = 13.8$, $df = 1,72$, $p < .01$. This compares well with the data from the original sample females: high-stability = 9.3, $N = 35$; low-stability = 5.0, $N = 35$; $F = 16.7$, $df = 1,68$, $p < .01$.

TABLE 2

*ACL Stability Scale Adjectives for Females**
Indicative Adjectives (19 Items)

Adjective	Proportion of High- Stability Females Endorsing Item	Proportion of Low- Stability Females Endorsing Item	Adjective	Proportion of High- Stability Females Endorsing Item	Proportion of Low- Stability Females Endorsing Item
active	86	58	initiative	57	31
alert	94	61	logical	83	53
appreciative	100	78	mature	74	33
efficient	86	50	outgoing	66	28
energetic	74	39	peaceable	86	58
fair-minded	94	72	precise	40	14
feminine	86	56	quick	43	17
forgiving	94	72	resourceful	49	22
generous	80	56	thorough	63	25
humorous	89	67			
<i>Contraindicative Adjectives (10 Items)</i>					
awkward	20	44	immature	17	47
conceited	06	31	inhibited	11	33
cynical	14	42	lazy	34	64
disorderly	09	31	moody	60	83
egotistical	11	36	preoccupied	11	36

* Differences significant $p < .05$.

A test-retest procedure, using an independent sample of 100 male and 100 female undergraduates, who were given the ACL twice under standard conditions, with an interval of three months, yielded acceptably high reliability coefficients for the *Stb* scale items: males = .81, females = .78. An interesting question which was addressed with this sample is whether it is sensible to try to develop a *Stb* scale, the reliable responses to which would predict instability (unreliability) in behavior. Consequently, the question whether *Stb* reliability is related to *Stb* score was examined by comparing the test-retest *Stb* reliability coefficients of high and low-stability Ss (the upper and lower quartiles of *Stb* scores). For the high-stability males ($N = 25$), the *Stb* reliability coefficient was .82, compared to .78 for the 25 low-stability males ($Z = 0.37$, $p > .05$). A similar comparison for female Ss yielded a *Stb* reliability coefficient of .80 for 25 high-stability vs. .76 for 25 low-stability females ($Z = 0.34$, $p > .05$). From this it can be concluded that the ACL *Stb* scale is equally reliable for high and low-stability Ss.

Because in the ACL format scale scores are correlated positively with total number of adjectives endorsed, a correction factor must be introduced. The procedure used by Gough and Heilbrun (c.f. Gough and Heilbrun, 1965) was followed identically in this study. Four-category standard score conversion tables were calculated for the ACL *Stb* scale, based upon a sample of 2,212 females and 2,805 males. Tables 3 and 4 present the standard score conversion data. Fundamentally, a high *Stb* scale score indicates a tendency to endorse ACL items in the direction of Ss who are consistent in self-description, while a low *Stb* scale score indicates a tendency to respond to the ACL items in a direction indicative of inconsistency in self-description.

Summarized in Table 5 are the statistically significant correlations between the *Stb* scale and other ACL scales as well as several widely used psychometric scales from the Minnesota Multiphasic Personality Inventory (MMPI). The *Stb* scale shows significant positive relations with those ACL scales which are associated typically with good personal adjustment, particularly in a college sample, e.g., Achievement, Self-Confidence, Endurance (c.f. Heilbrun, 1960, 1961 a, b). At the same time, it bears negative relationship to scales related to poor emotional adjustment, e.g., Succorance and Counseling Readiness of the ACL as well as several clinical scales of the MMPI.

Discussion

Inspection of the adjectives contained in Tables 1 and 2 shows that those items associated with high-stability describe very attractive, mature, and socially-desirable characteristics. On the other hand, the self-attributed qualities of low-stability Ss (*Stb* contraindicative items) are quite uncomplementary, self-critical, and socially-undesirable. In fact, the overlap between these groups of items and the ACL *Favorable Items* scale (*Fav*) and *Unfavorable Items* scale (*Unf*) is considerable. For male Ss, 23 of the 37 high-stability adjectives appear on the ACL *Fav*, while seven of the 18 low-stability items appear on the ACL *Unf*. Similarly, for female Ss, 11 of the 19 high-stability items are part of the ACL *Fav*, and six of the 10 low-stability adjectives are contained in the ACL *Unf*. Basically what this means, then, is that Ss who said generally favorable things about themselves on the first ACL administration

TABLE 3

Males—Conversion of Raw to Standard Scores for ACL Stability Scale

Raw	Total Number of Adjectives Endorsed			
	1-75	76-95	96-121	122-300
	Standard Scores			
-25		1	1	
-24	1	2	2	
-23	2	3	3	
-22	4	5	4	
-21	5	6	5	1
-20	7	7	7	2
-19	9	9	8	3
-18	10	10	9	4
-17	12	11	10	6
-16	13	13	11	7
-15	15	14	13	8
-14	17	15	14	9
-13	18	17	15	10
-12	20	18	16	12
-11	21	19	17	13
-10	23	21	19	14
-9	25	22	20	15
-8	26	23	21	17
-7	28	25	22	18
-6	30	26	23	19
-5	31	27	25	20
-4	33	29	26	21
-3	34	30	27	23
-2	36	31	28	24
-1	38	33	30	25
0	39	34	31	26
1	41	35	32	27
2	42	37	33	29
3	44	38	34	30
4	46	39	36	31
5	47	41	37	32
6	49	42	38	33
7	50	43	39	35
8	52	45	40	36
9	54	46	42	37
10	55	47	43	38
11	57	49	44	40
12	58	50	45	41
13	60	51	46	42
14	62	53	48	43
15	63	54	49	44
16	65	55	50	46
17	67	57	51	47
18	68	58	52	48
19	70	59	54	49
20	71	61	55	50
21	73	62	56	52

TABLE 3 (Continued)

Males—Conversion of Raw to Standard Scores for ACL Stability Scale

Raw	Total Number of Adjectives Endorsed			
	1-75	76-95	96-121	122-300
	Standard Scores			
22	75	63	57	53
23	76	65	58	54
24	78	66	60	55
25	79	67	61	56
26	81	69	62	58
27	83	70	63	59
28	84	71	64	60
29	86	73	66	61
30	87	74	67	63
31	89	75	68	64
32	91	77	69	65
33	92	78	70	66
34	94	79	72	67
35	95	81	73	69
36	97	82	74	70
37	99	83	75	71
38	100	85	76	72
39		86	78	73
40		87	79	75
41		89	80	76
42		90	81	77
43		91	82	78
44		93	84	79
45		94	85	81
46		95	86	82
47		97	87	83
48		98	89	84
49		99	90	86
50		100	91	87
51			92	88
52			93	89
53			95	90
54			96	92
55			97	93
56			98	94
57			99	95
58			100	96
59				98
60				99
61				100

tended to be quite stable in self-description on later ACL administrations, while those who described themselves unfavorably were much less stable in self-description over a period of several weeks.

TABLE 4

Females—Conversion of Raw to Standard Scores for ACL Stability Scale

Raw	Total Number of Adjectives Endorsed			
	1-78	79-98	99-119	120-300
Standard Scores				
-13			1	
-12	1		2	
-11	3		5	2
-10	6	2	7	4
-9	9	4	9	7
-8	12	7	12	9
-7	15	10	14	11
-6	18	13	16	14
-5	21	16	19	16
-4	24	18	21	18
-3	27	21	23	21
-2	29	24	26	23
-1	32	27	28	26
0	35	30	30	28
1	38	32	33	30
2	41	35	35	33
3	44	38	37	35
4	47	41	40	37
5	50	43	42	40
6	53	46	44	42
7	56	49	47	44
8	59	52	49	47
9	62	55	51	49
10	65	57	54	52
11	68	60	56	54
12	70	63	58	56
13	73	66	61	59
14	76	69	63	61
15	79	71	65	63
16	82	74	68	66
17	85	77	70	68
18	88	80	72	71
19	91	83	75	73
20	94	85	77	75
21	97	88	79	78
22	100	91	82	80
23		94	84	82
24		97	86	85
25		99	89	87
26		100	91	90
27			93	92
28			96	94
29			98	97
30			100	99
31				100

TABLE 5

Intercorrelations between Stability Scale and Other Measures^a

ACL Scales	r^b	ACL Scales	r
Defensiveness	58 ^c	Unfavorable	-56
Favorable	56	Succorance	-46
Self-Confidence	65	Abasement	-36
Self-Control	39	Counseling	
		Readiness	-41
Personal Adjustment	59		
Achievement	73	<i>MMPI Scales</i>	
Dominance	70	F	-31
Exhibition	26	D	-27
Endurance	51	D30 ^d	-40
Order	47	MAS	-36
Intracception	42	Pt	-36
Nurturance	38	Sc	-32
Affiliation	48	Si	-34
Heterosexuality	36	Pd	-23

^a Based on $N = 131$ undergraduates (66 males and 70 females).^b All correlations significant $p < .01$.^c Decimals omitted.^d From Dempsey (1964).

It is worth noting in this context that there are sex differences in ACL items that differentiate between high- and low-stability Ss. There are only five high-stability items for males and females in common; they are *alert, efficient, initiative, quick, and thorough*. For the low-stability adjectives, only three were common to both male and female Ss; they are *awkward, lazy, and moody*. Why there should be so few items in common across sex is not altogether clear, although a partial explanation may be that a considerable proportion of the remaining adjectives are related to sex-specific endorsement tendencies. For example, 10 of the 37 high-stability items for males are included as masculine items in the ACL Femininity scale (Parker, 1969).

In addition to these data and the obvious face-validity that the high-stability adjectives (Tables 1 and 2) have for indicating normal or good adjustment, there are further data to suggest a positive relationship between stability and adjustment. Using the ACL, Goodman and Mendelsohn (1969) recently surveyed psychotherapists nationally to obtain information about therapists' values and attitudes toward patients (males) they treat. Comparison of the male high-stability adjectives (Table 1) with therapists' percep-

tions of the adult male who has a satisfactory adaptation to himself and his environment (normal adult male), reveals a substantial overlap. Specifically, adjectives in common are *alert, clear-thinking, confident, conscientious, reasonable, reliable, responsible, stable, tactful, and tolerant*. Goodman and Mendelsohn provide no data on females for comparison. However, the clear suggestion from these data for males, and the overall ACL profiles for high- and low-stability Ss is that stability, as defined in this context, is related to level of personal adjustment.

It could be argued that a person who ascribes certain undesirable properties to himself, such as the low-stability characteristics *disorderly, egotistical, immature, lazy, moody, sulky, and worrying*, would, by definition, be suffering from personal problems or internal conflicts. In this sense, the results of this study are comparable to those obtained by Worell (1962; 1963), who found that college students with high "intraindividual conflict," as compared with less conflicted students, showed lesser stability in reaction time and discrimination responses. The fact that this relationship between stability and personality disturbance has been obtained with very different measures of stability and of maladjustment suggests that the phenomenon is a general one which merits further investigation.

Theoretically, these findings are in accordance with the views of Baumeister and Kellas (1968), who view low-stability as a "generalized expression of behavior pathology." Erikson's (1968) theoretical views may also be related to the present findings. Erikson suggests that young people who have not yet achieved a stable sense of identity will experience any of a variety of emotional difficulties. If stability in self-description may be taken as an index of identity confusion, Erikson's views are supported by the present data. The question of whether stability can always be taken as an indication of maladjustment remains to be answered. There are, of course, situations in which flexibility of response is adaptive. Future investigations might determine which specific types of stability are related to various measures of maladjustment, and which are not. Moreover, in addition to the adjustive implications of the *Stb* scale, it may contribute to the understanding of personality concomitants of behavioral dimensions such as rigidity of attitudes, interpersonal evaluations, or perceptual stability.

REFERENCES

- Baumeister, A. A. and Kellas, G. Intrasubject response variability in relation to intelligence. *Journal of Abnormal Psychology*, 1968, 73, 421-423.
- Berdie, R. F. Intra-individual variability and predictability. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1961, 21, 663-676.
- Block, J. Some reasons for the apparent inconsistency of personality. *Psychological Bulletin*, 1968, 70, 210-212.
- Carrigan, P. M. Intraindividual variability in schizophrenia: Unpublished doctoral dissertation, University of Michigan, 1963.
- Dempsey, P. A unidimensional depression scale for the MMPI. *Journal of Consulting Psychology*, 1964, 28, 364-370.
- Erikson, E. *Identity: Youth in Crisis*. New York: W. W. Norton and Co., 1968.
- Fiske, D. W. An intensive study of variability scores. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1957, 17, 453-465.
- Goodman, R. and Mendelsohn, G. A. Psychotherapeutic change and social adjustment: A report of a national survey of psychotherapists. *Journal of Abnormal Psychology*, 1969, 74, 164-172.
- Gough, H. C. and Heilbrun, A. B., Jr. *The adjective check list manual*. Palo Alto, California: Consulting Psychologists Press, 1965.
- Heilbrun, A. B., Jr. Personality differences between adjusted and maladjusted college students. *Journal of Applied Psychology*, 1960, 44, 341-346.
- Heilbrun, A. B., Jr. Client personality patterns, counselor dominance, and duration of counseling. *Psychological Reports*, 1961, 9, 15-25. (a)
- Heilbrun, A. B., Jr. Male and female personality correlates of early termination in counseling. *Journal of Counseling Psychology*, 1961, 8, 31-36. (b)
- Hull, C. L. *A behavior system*. New Haven: Yale University Press, 1952.
- Lentz, T. F. The reliability of opinionnaire techniques studied intensively by the retest method. *Journal of Social Psychology*, 1934, 5, 338-356.
- McNemar, Q. *Psychological statistics*. (3rd ed.). New York: Wiley and Sons, 1962.
- Mischel, W. *Personality and assessment*. New York: Wiley and Sons, 1968.
- Parker, G. V. C. Sex differences in self-description on the Adjective Check List. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 99-113.
- Worell, L. Response to conflict as determined by prior exposure to conflict. *Journal of Abnormal and Social Psychology*, 1962, 64, 438-445.
- Worell, L. Intraindividual instability and conflict. *Journal of Abnormal and Social Psychology*, 1963, 66, 480-488.

A ONE-STEP NOMOGRAPH FOR THE KOLMOGOROV-SMIRNOV TEST

M. REEB

Bar Ilan University, Israel

HERRICK (1969) has given a nomograph for finding the two-tailed Kolmogorov-Smirnov statistic D for $p = .05$. He uses a log transformation, plotting D as the vertical axis against the smaller sample size as the horizontal, for various ratios of the two sample sizes, from 1.0 (equal sample sizes) to 0 (one-sample case). The nomograph given below is more convenient to use in that no calculation at all, and only one setting of a ruler, are required, and the scales are simpler to use. More information is also obtained in that both .05 and .01 levels of significance are given and most cases of small sample size are covered.

Rationale

Two-Sample Case

The large-sample critical value of the K-S statistic D (Siegel 1956) is given for the two-tailed test by

$$D \geq k \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \quad (1)$$

where D is the largest difference between the two cumulated distribution functions of n_1 and n_2 cases at any one point, and k is a constant depending on the level of significance previously selected to reject the null hypothesis.

Now, in a right-angled triangle of lesser sides n_1 and n_2 , if a line be drawn bisecting the right angle to meet the hypotenuse, and its length is d , then

$$d = \sqrt{2} \frac{n_1 n_2}{n_1 + n_2} \quad (2)$$

Combining (1) and (2) we obtain

$$D \geq k \sqrt{\frac{\sqrt{2}}{d}} \quad (3)$$

If now we draw a nomograph with n_1 and n_2 as vertical and horizontal axes respectively then D , the required minimum significant difference, for a particular value of n_1 and n_2 , is a simple function of k and of d , the distance from the origin along the diagonal bisecting the right angle between the axes. The diagonal is, therefore, calibrated by calculating, for each D , the critical distance given, from (3), by

$$d = \frac{\sqrt{2} k^2}{D^2} \quad (4)$$

One-Sample Case

Here, as in Siegel (op cit),

$$D \geq \frac{k}{\sqrt{n}} \quad (5)$$

which corresponds to (1) when $n_1 = n$ and $n_2 \rightarrow \infty$. Thus in the nomograph the line "between n_1 and n_2 " becomes a line from n_1 parallel to the horizontal axis, intersecting the diagonal at D . Or else, using the same scheme as above,

$$D \geq \frac{k}{\sqrt{d'}}, \quad \text{and} \quad d' = \frac{k^2}{D^2} \quad (6)$$

where d' is measured along the vertical axis from the origin.

Construction of the Nomograph

In drawing the nomograph (Figure 1), the distances from the origin are calculated in terms of the required D s by means of equations (4) and (6). Values of k were calculated to four figures by interpolation from Smirnov (1948), and give for the two-sample case, $d = 2.6080/D^2$ for the .05 and $3.7482/D^2$ for the .01 levels of significance, and for the one-sample case, $d' = 1.8442/D^2$ for .05 and $2.6504/D^2$ for .01.

These functions were calculated at intervals of D of 0.01 for the

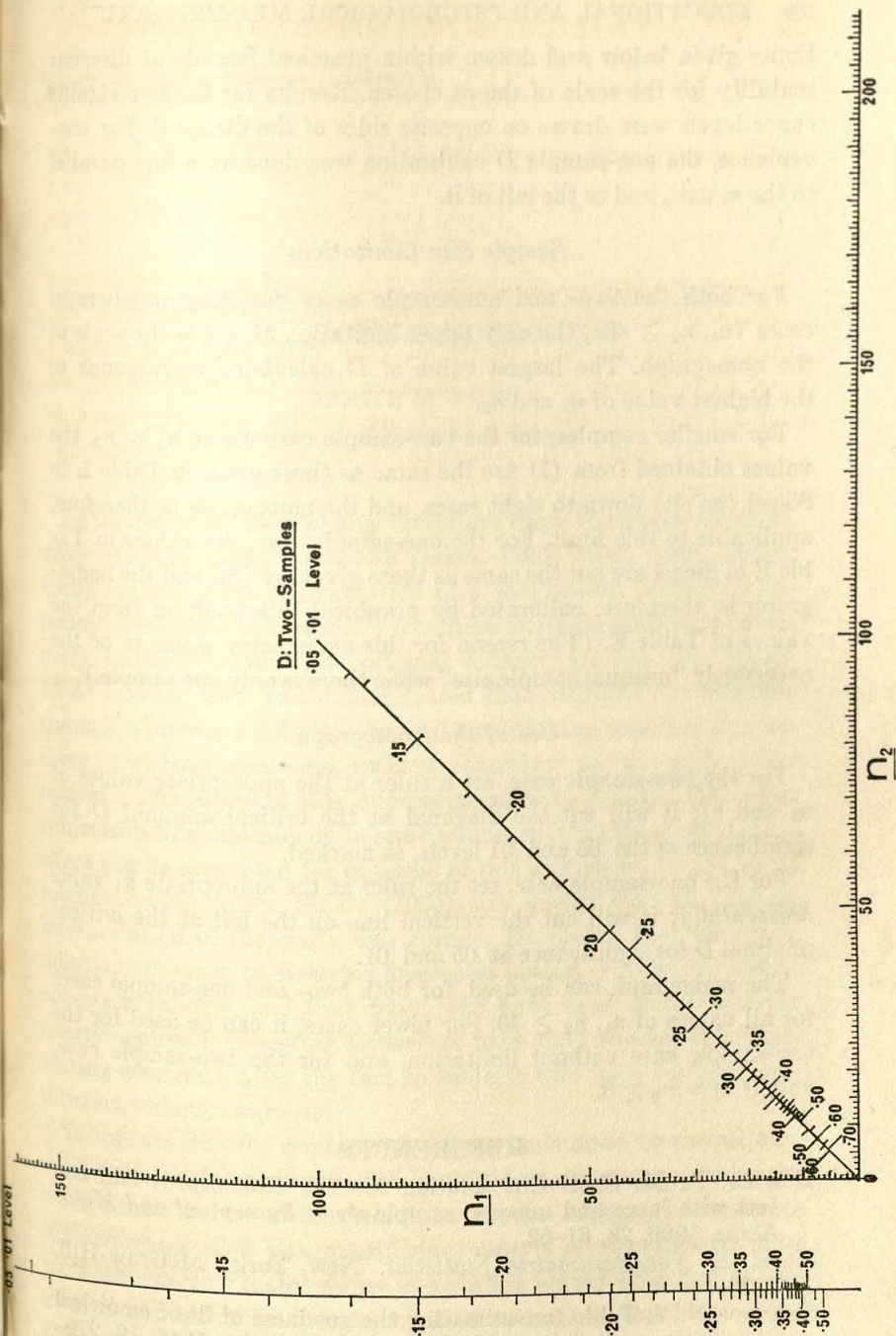


Figure 1. Nomograph for the Kolmogorov-Smirnov statistic D (two-tailed test).

limits given below and drawn within practical bounds of discriminability for the scale of the n s chosen. Results for the two significance levels were drawn on opposite sides of the diagonal. For convenience, the one-sample D calibration was done on a line parallel to the n_1 axis, and to the left of it.

Sample Size Limitations

For both the two- and one-sample cases, for large numbers of cases ($n_1, n_2 \geq 40$), the only upper limitation of size is the scale of the nomograph. The largest value of D calculated corresponds to the highest value of n_1 and n_2 .

For smaller samples, for the two-sample case where $n_1 = n_2$, the values obtained from (1) are the same as those given in Table L in Siegel (op cit) down to eight cases, and the nomograph is, therefore, applicable to this limit. For the one-sample case, the values in Table E of Siegel are not the same as those given by (5), and the nomograph is, therefore, calibrated by graphical interpolation from the values of Table E. (The reason for this asymmetry seems to be the necessarily "unequal sample size" when there is only one sample.)

Use of the Nomograph

For the two-sample case, set a ruler at the appropriate values of n_1 and n_2 ; it will cut the diagonal at the critical minimal D for significance at the .05 and .01 levels, as marked.

For the one-sample case, set the ruler at the appropriate n_1 value *horizontally*; it will cut the vertical line on the left at the critical minimal D for significance at .05 and .01.

The nomograph can be used, for both two- and one-sample cases for all values of $n_1, n_2 \geq 40$. For fewer cases, it can be used for the one-sample case without limitation, and for the two-sample case, *only if* $n_1 = n_2 \geq 8$.

REFERENCES

- Herrick, R. M. Short-Cut solution for the Kolmogorov-Smirnov test with large and unequal sample sizes. *Perceptual and Motor Skills*, 1969, 28, 61-62.
- Siegel, S. *Non-Parametric Statistics*. New York: McGraw-Hill, 1956.
- Smirnov, N. V. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 1948, 19, 279-281.

POSTEXPERIMENTAL ASSESSMENT OF AWARENESS IN ATTITUDE CONDITIONING

MONTE M. PAGE

University of Nebraska at Lincoln

In the recent attitude change literature there has arisen a controversy (Cohen, 1964; Insko and Oakes, 1966; Page, 1969; Staats, 1969) regarding the interpretation of a well known laboratory experiment intended to classically condition attitudinal affect to previously neutral stimuli. The original authors (Staats and Staats, 1958) claimed they had demonstrated that attitudes are acquired through a process similar to classical conditioning, and that this occurred "without awareness-without cognition" on the part of the subjects. The critics of this interpretation have asserted that some subjects in this situation do become aware and that the experimental effect can be accounted for in terms of this awareness. In his recent rebuttal to his critics, Staats (1969) still asserts that his original interpretation of the study was correct. He claims that his single question approach to assessing awareness postexperimentally was of adequate validity, and that those who have used more elaborate questionnaires in assessing awareness have only succeeded in suggesting awareness after the fact to subjects who were actually conditioned without awareness.

It appears that the crux of this debate boils down to the relative validity of single-question versus multiple-question awareness measures. Staats' assertion that multiple-question techniques of assessing awareness elicit postexperimental reports of awareness in subjects who weren't really aware is an often stated objection to the awareness position by those committed to a conditioning theory of both verbal operant and verbal classical conditioning experiments. This problem is an obvious possibility inherent in the methodological

necessity of assessing the awareness postexperimentally. However, there is evidence in verbal operant conditioning that the awareness occurs during and not after the conditioning procedure (De 1964; Page and Lumia, 1968). In evoking this argument regarding the present authors (Page, 1969) awareness results, Staats looked at an important factor. An attempt was made in that study to control for postexperimental suggestion by including in the postexperimental questionnaire items regarding the time at which subjects became aware. Subjects who did not clearly report that their awareness occurred during the experiment and prior to the marking of attitude scales were not counted as aware. In effect, Staats is suggesting that a large proportion of the subjects were dishonest regarding reports of the timing of their awareness. There is no evidence that this occurred. The fact that those and only those who accounted for the significant effect verbalized demand awareness supports the suggestion that suggestion by the interview procedure cannot account for the awareness. Why was awareness not suggested in the other study but was only reported by enough subjects to account for the conditioning effect and no more?

The purpose of the present study was to investigate data discrepancies between previous studies of the role of awareness of experimental demand characteristics (Orne, 1962) in attitude conditioning by comparing postexperimental assessment techniques. Staats and Staats (1957, 1958) asked a single open-ended question regarding subjects' thoughts about the purpose of the experiment. With a small percentage of subjects judged aware removed from the data, they found significance in the remaining data. In spite of a substantial literature in verbal operant conditioning (Spielberger and Nike, 1966) which demonstrates the inadequacy of a single open-ended question in detecting all the aware subjects, Staats still defends the validity of his technique for assessing awareness.

The basic argument against a single open-ended question as a measure of awareness is that, while the meaning of the question seems perfectly clear to its author, some subjects may misunderstand the import of the question. This is supported by the present authors' observations in previous studies that some subjects' answers to the first general and open-ended question are completely irrelevant to what is being asked. In addition, many subjects may understand the question, but write such a brief or vague answer as to preclude

accurate classification. For example, in previous studies some subjects have written simply "the purpose of the experiment was word association," then on the extended questionnaire gone on to explain that they meant there was an association of pleasant and unpleasant meaning for certain syllables. These subjects also stated that they had believed they were to demonstrate the learning of this by marking the syllables appropriately on the rating scales. Thus, the single question technique, while detecting some awareness, is not very accurate in its overall partitioning of awareness versus unawareness. It would seem that the procedure of removing aware subjects and analyzing the remaining data would require an accurate measure of awareness. With a single open-ended question this requirement is simply not met.

Insko and Oakes (1966) separated the concepts of contingency awareness, by which they meant knowledge of the consistent association between the affective words (beauty, sweet, pleasure, etc.) presented verbally and a specific nonsense syllable on the visually presented list, and demand awareness (Orne, 1962) or knowledge of the purpose of the experiment. Their measure of contingency awareness seems straightforward and adequate. One would expect this multiple-question technique to separate all of the contingency awares from the unawares. But, since no questions are included concerning the time and saliency of awareness, it might encourage guessing and attempts to recall events that were not very salient during the experiment. This assessment technique may identify more subjects as aware than actually were, but there is little danger that it would leave any awares undetected. Their 10-point scoring system, however, seems arbitrary and could lead to some ambiguity. For example, the only way a subject could receive a score of 1, 2, or 3 is to have made a number of incorrect guesses. There is no reason to believe that a subject who made incorrect guesses is any more unaware than a subject who did not guess and thus received the score of five. Likewise, at the other end of the scale there is no reason to believe that a subject who received a score of eight is any less aware than one who scores 10. Failure of a subject to be scored as aware on the first open-ended question may mean that he is less articulate but does not necessarily indicate that he was less aware of the contingency.

Insko and Oakes used a single open-ended question for assessing

demand awareness. The question was: "Did you feel as if you were supposed to rate the nonsense syllables in any particular way? If so explain." This seems to be a straightforward question, but it is subject to the same problems suggested regarding possible inaccuracies in Staats' technique for assessing awareness. What does it mean when a subject simply answers "no" to this question? It could mean that he did not interpret the question correctly or it could mean that he really was not demand aware. Since there are no other questions it is impossible to separate these two types of subjects. On the other hand, it would seem that a subject who says "yes" and then gives an incorrect explanation should be classified as clearly unaware rather than giving him some credit for being partially demand aware. The distinction between a subject who scores a two versus one who scores three seems also to lead to ambiguity. A subject who says "according to the way they were grouped" (Insko and Oakes, 1966) could be just as aware of the demand characteristics as one who gives a more specific answer, but since there are no more questions this could not be determined. What is being measured here is clarity of expression rather than demand awareness. There is thus the strong possibility that their measure and scoring procedure would lead to both false positives and false negatives, reducing the validity of the measure and hence, any correlation with conditioning.

The present author assumed that his more elaborate technique, utilizing multiple and converging questions for assessing demand awareness, correctly classified more subjects and was more valid. The author's measure did, in fact, account for all the significant conditioning effect (Page, 1969) while Insko and Oakes' (1966) measure of demand awareness did not. The basic issue in resolving these contradictory results is the question of which assessment technique is more valid. There were enough procedural differences between the original studies, however, so that the difference in results cannot be clearly attributed to questionnaire differences without further evidence. In the present study, the correlation between awareness and attitude conditioning was investigated as a function of the technique of assessment of awareness, in the context of the same study where all subjects were given all three questionnaires. To check on possible problems which may have arisen because of the necessity of not counterbalancing in the repeated measures design, the study was then repeated on three new groups of subjects where each group received only one questionnaire.

*Study I**Subjects*

Subjects were 160 introductory psychology students at the University of Nebraska at Lincoln. They were run in groups varying in size of from 15-30 each. Data were collected at approximately the middle of the university semester.

Method

All subjects were given the attitude conditioning procedure described in detail elsewhere (Page, 1969; Staats and Staats, 1957, 1958). There were four visually-presented nonsense syllables: wuh, yof, laj, and giw; and 18 conditioning trials. Wuh was paired with spoken words having strong evaluative meaning; the other three were paired with neutral words. Half ($N = 80$) of the subjects had positive evaluative meaning paired with wuh; the other half ($N = 80$) received negative meaning associated with wuh. Following this, subjects rated the four syllables (plus eight filler syllables not previously encountered) on 9-point pleasant-unpleasant scales of the semantic differential type. Then they circled words on a dittoed sheet which they remembered as having appeared on the spoken list.

Subjects then responded sequentially to three postexperimental questionnaires regarding awareness. First they were told to respond to the following question on the back of the sheet of paper used as the "second learning test": "Would you write down anything you thought about the experiment, especially anything you thought about the purpose of the experiment while you were participating in the experiment." This is the single open-ended question proposed by Staats (1969) and presumably something similar was used in his earlier studies. Papers were collected as subjects finished writing. When all papers were collected the next questionnaire was introduced by simply saying that the experimenter now wanted them to fill out a written questionnaire. Subjects then responded to this written version of the questionnaire used by Insko and Oakes (1966). Each question was on a separate page of a booklet and the booklets were collected as subjects finished the last page. Finally, subjects were given "honesty and conscientiousness" instructions (Page, 1968), and then asked to fill out a booklet containing the Page (1969) questionnaire. Added to this questionnaire as the sec-

and question were three 9-point rating scales asking subjects to rate their degree of attention, effort to learn all the words and boredom during the experiment. Otherwise the questionnaire was identical to that used previously.

The questionnaires were always presented in the above order. This was considered a necessity, though it confounds effectiveness of the questionnaire with order of presentation, because the simpler open-ended questionnaires would not be comparable to earlier results if they followed the more detailed and specific questionnaire. The possibility of the open-ended questions influencing the more detailed questionnaire was thought to be less important. Study II repeats the comparison of assessment techniques using a separate groups design which precludes the problem of one assessment technique influencing the other. A repeated measures approach was taken in this first study because the author wanted to explore cases where the same subject would be classified differently on the different measures of awareness. A direct comparison of the three assessment techniques on the same subjects seemed to be the best way of exploring differences between techniques.

All questionnaires were scored blind by two judges working independently, according to the rules employed in the previous studies where each questionnaire was used.

Results

The first question concerns the reliability of scoring by the two independent judges on the various measures of awareness. Since awareness was conceived of as a functional dichotomy, each judge's scoring was dichotomized and reliability was measured in terms of phi coefficients. For the Staats question, data were already in the aware-unaware form. For a subject to be classified aware on this question he had to state that wuh was associated with words of pleasant meaning (or unpleasant, depending upon the condition), otherwise he was scored unaware. This is essentially a measure of contingency awareness. Since few subjects clearly stated that they thought they were supposed to rate the syllable according to the association, no scoring for demand awareness was possible.

For the Insko and Oakes contingency awareness measure, the data were dichotomized by considering a score of six or below as unaware and seven or above as aware. On the demand awareness

measure the data were dichotomized between a score of one and two. For the Page measures of awareness, the data were dichotomized between a score of two and three on the 4-point scale of definitely unaware to definitely aware.

Table 1 presents the resulting phi coefficients (over their appropriate phi-max values) for scorer reliability. Below each phi over phi-max is presented, in parentheses, the ratio of phi over phi-max which will aid the reader in comparing the relative strengths of the various associations. It may be seen that scorer reliability is no problem for the Staats questionnaire. The judges agreed almost perfectly as to how all subjects should be scored. The dichotomous scoring of the Insko and Oakes questionnaires results in lower reliabilities than reported in the original study. While the judges seldom disagreed more than one scale score, it happened that the cut-off points selected were the ones where judges disagreed the most. Particularly on demand awareness, the judges had the most difficulty agreeing on whether a subject's explanation of his "yes" was incorrect or partially correct. When Pearsonian correlations were computed on the full scale scores, the reliability for contingency awareness was $r = .93$ and for demand awareness was $r = .89$. These figures are more comparable with the reliabilities reported by Insko and Oakes. It may be seen that the reliabilities for the Page awareness measures are adequate but not especially high. This is probably due to a failure on the part of the author to adequately train the other judge on the specific criteria for judging a subject to be aware. More of the disagreements were in the direction of the other judge attributing awareness where the author did not. This situation was corrected prior to the scoring of Study II, and as will be reported, the reliabilities were much higher in that study.

A more important question concerns the correlation of the awareness measures with the conditioning behavior. Table 1 also presents these phi correlations. The awareness dichotomies used here were based on the average of both judges' ratings. The 9-point rating scale (conditioning) was dichotomized by considering a subject in the positive condition who scored one or two to show conditioning, the others were considered not to have shown conditioning; this fit the bimodality of the data. This scoring was reversed in the negative condition, and then the data were pooled for the correlations. It may be seen that while the Staats questionnaire can be reliably scored,

TABLE 1

Reliability and Validity of Three Postexperimental Measures of Contingency and Demand Awareness in Study 1 Expressed as Phi over Phi-max

Correlations	Staats Contingency awareness	Questionnaires		Page	
		Insko & Oakes Contingency awareness	Demand awareness	Contingency awareness	Demand awareness
Scorer	.88/.90	.84/.92	.71/.97	.72/.83	.80/.90
reliability	(.98)*	(.91)	(.73)	(.87)	(.81)
Correlation with	.34/.63	.57/.90	.41/.71	.57/.87	.60/.87
criterion con- ditioning	(.54)	(.63)	(.57)	(.65)	(.84)

* Ratio of phi over phi-max.

it does not correlate especially high with conditioning. This could be, as Staats claims, due to the fact that his open-ended question does not suggest as much awareness to genuinely conditioned subjects. But, in the context of the rest of the data it is more likely that the open-ended question is simply not a very valid measure of awareness.

The Insko and Oakes, and the Page measures of contingency awareness correlate with conditioning to the same degree, and at about the same levels reported in the original studies. A discrepancy between the measures of demand awareness also occurs, as would be expected, from the correlations reported in the original studies. The Page measure correlates much better with conditioning than does contingency awareness, while the Insko and Oakes measure does not correlate as well as contingency awareness. The dichotomized scoring used in this analysis actually makes the Insko and Oakes measure a little more valid than the original scoring method. The Pearsonian correlation using the full scale scores for both demand awareness and conditioning was $r = +.34$, which is more comparable to the original study.

It is possible to attribute this discrepancy between the Page, and Insko and Oakes measures to a lack of validity in the Insko and Oakes open-ended question approach. Examination of the *Ns* in Table 2 reveals that the Insko and Oakes measure identifies fewer subjects as aware. What is not apparent in this table is that the discrepancy is even greater than this, because the Insko and Oakes measure also identifies several subjects as demand aware which were not later demand aware on the more elaborate Page question-

naire. The subjects aware on the Page measure, but not on the Insko and Oakes measure, in general show high conditioning (14 of 16). The ones aware on the Insko and Oakes measure, but not on the Page measure, in general do not show it (1 of 7). Thus, it is possible to suggest that the Insko and Oakes measure misclassifies a number of subjects or is of low validity. Therefore, it would not be expected to account for all the variance in attitude conditioning, even if, as the author has suggested, demand awareness is the crucial variable in so-called attitude conditioning.

It is the usual practice in both verbal operant and classical conditioning studies to remove aware subjects from the data before analysis. Such a procedure implicitly assumes a valid measure of awareness. Table 2 shows what happened to the conditioning effect when subjects aware by the various measures used in this study are removed before analysis. Notice first that the total data showed a strong conditioning effect ($t = 5.19, p < .001$). The Staats question identifies 19 of 160 subjects as contingency aware. Four of these were contingency, but not demand aware on the Page measures, and they did not show conditioning. When the Staats awares are removed the means are less discrepant, but the difference is still highly significant ($t = 3.83, p < .001$). This is exactly what would be expected if the validity of the measure was not adequate and many awares were left in the data. Notice that the Insko and Oakes contingency awareness measure identifies 60 of 160 subjects as aware, and removal of these from the data leaves the remaining

TABLE 2

Mean Conditioning with Awares Removed by Various Measures of Awareness

Subjects	Direction of conditioning	
	Positive	Negative
Total data—none removed	$\bar{X} = 3.58$ $N = 80$	$\bar{X} = 5.58$ $N = 80$
Staats' CA removed	$\bar{X} = 3.86$ $N = 71$	$\bar{X} = 5.39$ $N = 70$
Insko & Oakes CA removed	$\bar{X} = 4.51$ $N = 47$	$\bar{X} = 4.70$ $N = 53$
Page CA removed	$\bar{X} = 4.23$ $N = 56$	$\bar{X} = 4.61$ $N = 54$
Insko & Oakes DA removed	$\bar{X} = 3.76$ $N = 67$	$\bar{X} = 4.98$ $N = 64$
Page DA removed	$\bar{X} = 4.19$ $N = 63$	$\bar{X} = 4.47$ $N = 59$

data not significant ($t = .44$). According to the author's theory (Page, 1969) of the role of demand awareness in this situation a good measure of contingency awareness should do just that, because demand awareness requires that a subject first be contingency aware. Notice next that the Page measure of contingency awareness also renders the remaining data nonsignificant ($t = .95$) by removing 50 subjects from the data.

The crucial importance of demand awareness as measured by the Page questionnaire is illustrated by the fact that it also renders the data nonsignificant ($t = .72$), but it does so by removing only 38 of 160 subjects. That is, it approaches the ideal of removing those and only those who account for the significant effect, and this accounts for the higher correlation reported earlier. This may be taken as very strong support for the demand characteristics position.

As previously suggested, the Insko and Oakes measure of demand awareness seems to lack validity. It removes 29 of 160 subjects from the data, but the remaining data are still highly significant ($t = 3.05$, $p < .01$). In fact, some of the subjects ($N = 5$) identified as demand aware by this measure are not even contingency aware by any measure; this would be a logical impossibility if the measure were valid. On the other hand, the measure leaves a number of subjects ($N = 14$) who are contingency aware and high conditioners in the data. It is precisely these subjects that are identified as demand aware by the Page questionnaire.

Recall that subjects were asked to rate their attention, effort to learn and boredom on 9-point rating scales. Subjects were divided into demand aware versus unaware groups on the basis of the Page measure. Table 3 presents these data. The aware subjects reported more attention ($t = 2.72$, $p < .01$), more effort to learn all the

TABLE 3

Mean Reported Attention, Effort to Learn and Boredom for Demand Aware ($N = 58$) versus Unaware ($N = 122$) Subjects

Item	Aware	Unaware
Frequently did not pay attention, to tried very hard to pay attention	6.92	5.94
Did not try my best to learn, to tried to learn all the words	6.77	5.74
I was not especially bored, to I was very bored	3.20	4.33

words ($t = 2.94, p < .01$) and less boredom ($t = 2.76, p < .01$) during the experiment than the unawares. While these data do not distinguish between the predictions of conditioning versus demand awareness theory, they do suggest that attentional and motivational variables are important in distinguishing between subjects who are aware and show the conditioning effect and those who do not.

Study II

Subjects

Subjects were 300 introductory psychology students at the University of Nebraska at Lincoln. They were run in groups varying in size from 15-30 each. Data were collected slightly past the middle of the university semester.

Method

The method and procedure was the same as in Study I with the following exceptions: (a) since the data were to be used only for correlational purposes, all subjects were run in the negative condition rather than reversing conditions for half the subjects; (b) subjects were divided into groups of 100 and each group was given a different postexperimental questionnaire for awareness rather than all subjects receiving all questionnaires. If the pattern of correlations is similar to that found in Study I, then this would be evidence that the former correlations could not be accounted for by the order in which the questionnaires were presented; (c) due to the lower than expected scorer reliabilities obtained on the Insko and Oakes, and Page questionnaires in Study I, the judges spent some time discussing in detail the scoring criteria for all measures. Particularly the protocols for subjects where the judges had disagreed in Study I were discussed in detail. It was expected that this would increase scorer reliability in this study. The judges then scored the questionnaires from Study II independently.

Results

The results of this study are presented in Table 4. Comparison of this table with Table 1 reveals that the scorer reliability on all measures is higher in this study. With the exception of the Insko and Oakes demand awareness measure, these reliabilities are remarka-

bly high, demonstrating at least that judges agree more after they have practiced and discussed than before.

Inspection of the pattern of correlations between dichotomized awareness measures and the conditioning scale dichotomized at eight and nine (high conditioning) versus one through seven (low conditioning), reveals a pattern similar to that found in Study I. The poorest correlate with conditioning is the Staats open-ended question and the strongest is the Page measure of demand awareness. Since this was a separate groups design, and since the correlations are approximately the same as in the earlier repeated measures design, it may be concluded that order of presentation probably did not have an important effect on the outcome of Study I. Since this replication again finds demand awareness during the experiment as measured by the Page questionnaire to be the strongest correlate with conditioning, it again points to the crucial importance of this variable.

Discussion

These studies have brought evidence to bear on the questions raised earlier (Page, 1969) concerning the nature of demand awareness and the appropriate method of assessing it postexperimentally. It appears clear now that the Insko and Oakes (1966) approach to scoring responses to a single open-ended question is more of a measure of clarity of expression in responding to the question as well as of subjects' understanding of what is being asked than it is a measure of demand awareness. These variables are correlated but not strongly. What is required is a multiple question approach, so that

TABLE 4

Reliability and Validity of Three Postexperimental Measures of Contingency and Demand Awareness in Study II Expressed as Phi over Phi-max

Correlations	Questionnaires				Page Demand awareness
	Staats Contingency awareness	Insko & Oakes Contingency awareness	Demand awareness	Contingency awareness	
Scorer	.76/.76	.92/.96	.78/.98	.89/.93	.90
reliability	(1.00)*	(.96)	(.79)	(.96)	(.73)
Correlation with	.26/.61	.52/.84	.53/.88	.59/.87	.73
criterion con-	(.42)	(.62)	(.60)	(.67)	(.67)
ditioning					

* Ratio of phi over phi-max.

there is less possibility of a breakdown in communication. The subject has to understand what is being asked and he has to write enough so that the judge can understand what he meant.

There are numerous studies in the current literature which either conclude that conditioning can occur without awareness because unawares by their measure show a significant effect or that conditioning is a function of awareness because unawares by their measure showed no effect. It seems that the resolution of these contradictory conclusions is basically a measurement problem. Attention should not be focused on the contradictory conclusions but on the differences in measurement operations. The logic of classifying subjects as to awareness, removing the awares from the data and then drawing conclusions about the remaining data requires accurate classification. The present studies suggest that brief open-ended measures of awareness are simply not valid, and this can account for the current proliferation of contradictory data. An interesting fact, often overlooked because attention is focused upon the presence or absence of conditioning in subjects classified as unaware, is that aware subjects by any measure always show much more conditioning than unawares. Thus, there is the strong possibility that only a few awares left in the data because of an invalid measure would result in "conditioning without awareness," while a more valid measure of awareness would not.

More generally, one often encounters deception experiments in the contemporary social psychology literature where subjects were asked an open-ended question concerning subjects' knowledge of the purpose at the conclusion of the experiment. The authors usually indicate that a certain small percentage of subjects saw through the deception and that these were eliminated prior to the analysis. But, we now ask, how many others also were aware and were not detected? What effect might these undetected demand awares have on the significance of the data? In the light of the present data, it is strongly recommended that we develop appropriate extended post-experimental questionnaires and use them whenever conducting a deception experiment. Especially this should be done with experimental situations where in the past a few awares have been detected using the single question approach. Whenever a few verbalized awareness to an open-ended question there may have been several more who would have been detected by a more valid technique.

It is often warned that the use of extended awareness interviews can lead to the suggestion of verbalizations of awareness after the fact. The present author is well aware of this possibility, therefore controls for this were built into the Page (1969) questionnaire. It should be pointed out, however, that this excuse for not using an adequate questionnaire has been overworked, particularly by those with some theoretical investment in not discovering that, after all, subjects in human experiments do think about the purpose of what they are required to do. The author feels that it is difficult to suggest something as basic as whether the subject knew what was going on in the experiment or not, especially with a more elaborate questionnaire. With a single open-ended question the subject's answer may be so vague that it may suggest to the scorer that he might have been aware when in fact he was not. Rather than suggesting awareness, more elaborate questionnaires actually reduce the possibility of the judge reading into the subject's reports.

There is evidence in this study, however, that some subjects can reflect back on their previous experience and verbalize the correct contingency when, in fact, it was not very salient during the experiment, at least when there are no controls in the questionnaires to prevent this. The Insko and Oakes contingency awareness measure is especially vulnerable to this, because subjects are asked to verbalize the contingency in any way they can, disregarding time or saliency of awareness. Recall that in Study I the Insko and Oakes contingency awareness questionnaire identified 10 more subjects as aware than did the Page measure of contingency awareness. The Page measure considered both verbalizations of the contingency and a rather stringent criterion of when the awareness occurred. These 10 subjects were indeed ones who on the Page questionnaire said, in effect, "Yes, I recall some bad words associated with wuh but I'm not really sure this was always the case and I didn't think much about it until afterwards." If, indeed, it is demand awareness during the experiment that mediates attitude conditioning, then subjects contingency aware after the fact would not be expected to show conditioning and, in fact, they do not. A high correlation between awareness and behavior depends upon accurate assessment of awareness. Subjects who are aware but missed by an insensitive assessment technique as well as subjects unaware but suggested aware by a probing technique both serve to reduce correlations. It is here

recommended that a solution to this problem is to use a more extended and probing questionnaire so as to avoid missing aware subjects, and to include questions concerning the timing and saliency of awareness so as to avoid suggesting reports of awareness in unaware subjects.

There is the remote possibility of suggesting awareness in some subjects even with a controlled questionnaire; however, it seems scientifically more rigorous and defensible to risk suggesting awareness to a few subjects who really were not, than to risk not identifying subjects who actually were aware if, as is usually the case, one is concerned with the extent of the experimental effect in subjects who are truly unaware.

Summary

Two studies compared three postexperimental techniques for assessing awareness in attitude conditioning. It was found that multiple question techniques for assessing both contingency and demand awareness resulted in stronger correlations with conditioning than did single open-ended question techniques. This was attributed to ambiguities inherent in the open-ended technique which leads to misclassification of many subjects as to their awareness. The possibility of suggesting awareness by using extended questionnaires was discussed, and it was concluded that questions regarding the time and saliency of subjects' awareness should be included in questionnaires.

REFERENCES

- Cohen, B. H. Role of awareness in meaning established by classical conditioning. *Journal of Experimental Psychology*, 1964, 67, 373-378.
- De Nike, L. D. The temporal relationship between awareness and performance in verbal conditioning. *Journal of Experimental Psychology*, 1964, 68, 521-529.
- Insko, C. A. and Oakes, W. F. Awareness and the "conditioning" of attitudes. *Journal of Personality and Social Psychology*, 1966, 4, 487-496.
- Orne, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 1962, 17, 776-783.
- Page, M. M. Modification of figure-ground perception as a function of awareness of demand characteristics. *Journal of Personality and Social Psychology*, 1968, 9, 59-66.
- Page, M. M. Social psychology of a classical conditioning of atti-

tudes experiment. *Journal of Personality and Social Psychology*, 1969, 11, 177-186.

Page, M. M. and Lumia, A. R. Cooperation with demand characteristics and the bimodal distribution of verbal conditioning data. *Psychonomic Science*, 1968, 12, 243-244.

Spielberger, C. D. and De Nike, L. D. Descriptive behaviorism versus cognitive theory in verbal operant conditioning. *Psychological Review*, 1966, 73, 306-326.

Staats, A. W. Experimental demand characteristics and the classical conditioning of attitudes. *Journal of Personality & Social Psychology*, 1969, 11, 187-192.

Staats, A. W. and Staats, C. K. Attitudes established by classical conditioning. *Journal of Abnormal and Social Psychology*, 1958, 57, 37-40.

Staats, C. K. and Staats, A. W. Meaning established by classical conditioning. *Journal of Experimental Psychology* 1957, 54, 74-80.

A PROJECTIVE OCCUPATIONAL ATTITUDES TEST¹

LEROY C. OLSEN

AND

WILLIAM H. VENEMA

Washington State University

OCCUPATIONAL counseling is a personalized process which attempts to aid the individual in understanding his values as they relate to occupational selfactualization. Goodstein (1965) suggested that an adequate theory of vocational development, choice, and adjustment must take into consideration both external reality factors and psychodynamics. Forer (1965) suggested that a comprehensive occupational theory must account for processes and problems of developing skills, knowledge, efficiency, productivity, creativity, attitudes, and interpersonal relationships.

Means of assessing occupational attitudes are limited in nature. Frequently used standardized paper-and-pencil interest tests such as the Kuder Preference Record and the Strong Vocational Interest Blank may only superficially consider attitudes. Allport (1954) believed that such instruments were helpful to a degree, but they dealt with the typical variables, and thus provided very little information about the unique motivation or the underlying potentialities of a single case.

While the results of research to date do not appear to have produced significant results in all cases, certain facts seem apparent. First, most of the research has been conducted with college students or professional groups; second, standard projective devices have been used rather than adapting or developing devices utilizing

¹This research was supported in part by the Office of Education, U. S. Department of Health, Education, and Welfare, Project No. ERD-257-65 and Project No. OE7-0031.

occupational themes; and third, projective devices have been used primarily for the purpose of studying personality rather than attitudes.

Purpose of the Study

The purpose of this study was an attempt to validate and standardize a projective device for measuring the attitudes of culturally disadvantaged youth toward work roles and work environments. It was assumed that the following aspects of attitudes could be measured by a projective occupational attitude technique:

1. That the projective technique would provide an index of attitudes towards the following occupational aspects; tasks, tools, equipment, working environment, and interpersonal relationships representative of certain occupations;
2. That the projective test would provide meaningful attitudinal information that could be classified, validated, and would distinguish between various groups; and
3. That the projective test would provide meaningful attitudinal information in relation to self-concept, needs, and occupational choices.

The study also attempted to obtain data on the following attitudes: Attitudes toward satisfaction with parental relationships (a) occupational status and (b) socioeconomic status; long-range occupational goals; occupational (a) security, (b) prestige, (c) achievement, and (d) satisfaction-enjoyment; plans for occupational advancement; (a) immediate economic press and (b) delayed need gratification; and anxiety and frustration in (a) meeting job qualifications and (b) gaining entrance and employment in need satisfying occupations. Attitudes expressed toward possible interpersonal relationships with other workers; through identification with the workers; toward supervisory or authority figures; and toward (a) tasks, (b) tools, (c) equipment, and (d) working conditions.

Procedure

Sample

One portion of the sample was obtained from the Columbia Job Corps Center located at Moses Lake, Washington. Youths selected

for the Job Corps included those who: (a) were school dropouts, (b) were unable to find or hold jobs or lack marketable skills or expressed vocational goals, (c) had poor school performance, (d) were unable to pass the educational part of the Selective Service Examination, and (e) had a self-concept of defeat and failure. A total of 88 enrollees met the criteria for selection and constituted the study sample. Of the 88, a total of 47 or 53.41 per cent were Negro and 41 or 46.59 per cent were Caucasian. From each of the Negro and Caucasian groups a random sample of 15 subjects was selected for the study.

A secondary school and two junior high schools provided a sample of 384 pupils from the Tacoma Public Schools, Tacoma, Washington. Subjects were selected on the basis of variables similar to those necessary for persons enrolling in the Job Corps. Of this total, 91 or 23.35 per cent met the study criteria. From each of the Negro and Caucasian groups a random sample of 15 subjects was selected for the study. Two groups were selected from the secondary school, and two groups from each of the two junior high schools. The total sample included 90 subjects.

Measuring Instrument

Olsen (1966) developed a projective instrument, Projective Occupational Attitudes Test (P.O.A.T.) designed to assess occupational attitudes of non-college bound pupils. The test consists of 10 pictures mounted on heavy paper portraying five major dimensions of common nonprofessional level male occupational situations, and one blank card. The dimensions portrayed are (a) acts, (b) tools and/or equipment, (c) materials, (d) working environments, and (e) interpersonal relationships. The test attempts to provide a measure of occupational attitudes of noncollege bound individuals in the following occupational areas: Distribution, Carpentry, Electrical, Store Clerk, Farm Work, Heavy Construction, Forestry, Traffic, Janitorial Work, Service Station, and a Blank Card. It should be recognized that this study was exploratory and represented an original instrument and the initial research with this instrument. Complete scoring details are contained in the original report.

Statistical Treatment of Data

Intelligence test standard deviation scores, reading grade level scores, family stability, occupational models, and occupational status of subjects' parents were used for classifying subjects into various categories. The Fisher Exact Probability Test and the Chi Square test were applied to determine the significance of proportions of subjects falling into various categories between groups. In situations appropriate for testing the level of significance of difference between two means the *t* test was applied.

Results and Discussion

Seven variables were considered under total response: perception and response, movement in story, response to the total card situation as compared to some aspect, a simple description as compared to an organized story, acceptance as compared to rejection of the stimulus card, projection—degree and kind, and indications of liking or disliking the work pictured on the cards.

There were no significant differences between groups in proportions of subjects falling into the various categories on the following variables related to total response: movement in story, acceptance as compared to rejection of the stimulus card, projection—degree and kind, and liking compared to disliking the portrayed work.

Senior high Negroes gave significantly more adequate responses than senior high Caucasians, .01 level; junior high Negroes, .01 level; and junior high Caucasians, .05 level. Total Job Corps enrollees gave significantly more adequate responses than total junior high pupils, .05 level. Total senior high pupils gave significantly more adequate responses than total junior high pupils, .05 level.

Senior high Negroes responded significantly more frequently to the total card situation than senior high Caucasians, .01 level; Job Corps Caucasians, .05 level; and junior high Caucasians, .05 level. Total senior high Caucasians responded significantly more frequently to the total card situation than total junior high school pupils, .02 level.

The majority of subjects' total responses, with the exception of senior high Negroes and total senior high pupils, were classified as minimal or inadequate in perception and response to the stimulus cards. Most subjects gave simple descriptions rather than organized

responses. Movement was usually included in these responses. The subjects appeared to view being employed in any occupation as being more desirable than being unemployed or identifying with a prestigious occupation.

The results for physical aspects of work indicated a low level of concern about tools and equipment. The subjects responded more frequently to tasks and interpersonal relationships than tools and equipment.

Only two of the 19 comparisons between groups on attitudes toward equipment were significant. Senior high Negroes were significantly more concerned with equipment than senior high Caucasians, .02 level, and junior high Negroes, .05 level. Those subjects who did respond to tools and equipment were usually unable to accurately describe the appropriate uses of the tools and equipment.

The subjects were more concerned with tasks than the other three variables. Only one of the comparisons between groups for tasks was significant. Junior high Caucasians were significantly more concerned with tasks, .05 level, than senior high Caucasians.

Senior high Negroes were significantly more concerned with work environment than junior high Negroes, .02 level, and junior high Caucasians, .05 level. Total Job Corps enrollees were significantly more concerned with work environment, .05 level, than total junior high pupils. Total senior high pupils were significantly more concerned with work environment, .05 level, than total junior high pupils.

The aspects of work environment were related to the placement of figures in identifiable surroundings. Senior high Negroes preferred occupations performed outside, while other groups did not indicate a definite preference for occupations performed either outside or inside. Junior high Negroes were least concerned with work environment than other groups.

In relation to other variables, security was the chief concern of all groups. Senior high Caucasians were more concerned with security, .05 level, than Job Corps Negroes. Total junior high pupils were significantly more concerned with security, .05 level, than total Job Corps enrollees. Total senior high pupils were significantly more concerned with security, .05 level, than total junior high pupils.

The need for satisfaction and enjoyment in work was ranked second by all groups. The satisfaction and enjoyment variable was concerned with a sense of pleasure or gratification obtained from work. The results indicated that senior high Negroes were significantly more concerned, .01 level, than Job Corps Negroes. Senior high Caucasians were significantly more concerned, .05 level, than Job Corps Negroes. Total senior high pupils were significantly more concerned, .02 level, than total junior high pupils. Job Corps Negroes were less concerned about satisfaction and enjoyment in work than other groups. Total senior high pupils tended to identify at higher occupational levels than other groups.

The achievement variable tended to differentiate between groups more frequently than any of the POAT variables. The achievement variable was concerned with such factors as going to school, getting a job, work experience, and occupational advancement. Ten of the 19 comparisons between groups were significant. The results indicated that senior high Negroes were significantly more concerned with achievement than Job Corps Negroes, Job Corps Caucasians, and junior high Negroes, .01 level; and junior high Caucasians, .05 level. Senior high Caucasians were significantly more concerned with achievement than Job Corps Negroes and Job Corps enrollees, .05 level. Total senior high pupils were significantly more concerned, .01 level, than total junior high pupils. Total junior high pupils were significantly more concerned with achievement, .05 level, than total Job Corps enrollees. Job Corps Negroes, Job Corps Caucasians, junior high Negroes, and junior high Caucasians were least concerned with achievement. There were no significant differences between groups in the proportions of subjects falling at either the mean or above category, as compared to the below-the-mean category for security-money. The need for money or the concern over money was ranked fourth by the total group.

Senior high Negroes were significantly more concerned with prestige than junior high Caucasians, .01 level; junior high Negroes, .01 level; and senior high Caucasians, .05 level. Total senior high pupils were significantly more concerned with prestige than total junior high pupils, .01 level, and total Job Corps enrollees, .05 level.

Senior high Negroes were significantly more concerned with interpersonal relationships, .02 level, than junior high Negroes. Total

senior high pupils were significantly more concerned with interpersonal relationships, .02 level, than total junior high pupils.

Junior high Negroes were less concerned over interpersonal relationships than other groups. They also tended to identify at lower occupational levels. There were no significant differences between groups for dependency. In general, there tended to be a low concern over dependency by all groups.

The results for the POAT primary occupational identification levels tended to reflect the earlier results of the subjects' selections of the POAT Blank Card occupational identification levels. Senior high Negroes identified at significantly higher occupational levels than Job Corps Negroes, .01 level; junior high Negroes, .01 level, Job Corps Caucasians, .05 level; and junior high Caucasians, .05 level. Senior high Caucasians identified at significantly higher occupational levels, .02 level, than junior high Negroes. Total senior high pupils identified at significantly higher occupational levels than total Job Corps enrollees, .01 level, and total junior high pupils, .01 level.

The results were similar to those found in subjects' selections of Blank Card primary occupational identification levels. The results also indicated that senior high Negroes aspire to or identify with occupations approximating the semiprofessional levels (Roe). Junior high Negroes identified at lower levels than all other groups.

There were no significant differences between groups for acceptance or rejection of the primary identification figures. As a total group the majority of subjects tended to accept the primary occupational identification figures. There were no significant differences between groups for favorable authority relationships compared to unfavorable authority relationships. Junior high Negroes tended to be slightly more resentful of authority figures than other groups.

Senior high Negroes utilized defense mechanisms in their responses significantly more frequently than Job Corps Negroes, Job Corps Caucasians, junior high Negroes, junior high Caucasians, and senior high Caucasians, .01 level. Total senior high pupils utilized defense mechanisms significantly more frequently than total junior high pupils.

The results suggested a lack of parental influence in the subjects' perceptions and attitudes toward work. There were no sig-

nificant differences between groups concerning attitudes and feelings toward parents. Inspection of the data revealed that only 96 responses relating to parents were made by the 90 subjects.

No significant differences were found between groups for the two variables relating to long range occupational goals. The importance of these results is evident because each subject was specifically instructed to respond by indicating what the persons portrayed in the stimulus cards felt the future held for them. It seems that the press for security and immediate need gratification fosters attitudes which are probably related to immediate employment and need gratification, rather than to long range occupational planning or interest.

One significant difference was found in the comparisons involving occupational advancement, Job Corps Caucasians indicated more frequently that the portrayed figures on the cards would move upward in occupations than Job Corps Negroes; this was significant at the .02 level. The majority of the subjects felt that the portrayed figures would be static or remain at their present occupational level.

The results indicated no significant differences between groups for concern over ability and aptitude in meeting job qualifications in need satisfying occupations. The results suggested a limited perception by subjects of their aptitudes and ability as these variables relate to employment.

Senior high Caucasians were significantly more concerned over possible denial of training, skill, and education, .05 level, than junior high Negroes. Total senior high pupils were more concerned with possible denial of training, skill, and education, .001 level, than total junior high pupils.

As a total, the subjects responded more frequently to possible denial of training, skill, and education than to any of the variables related to anxiety and frustration in meeting job qualifications; the one exception was junior high Negroes. Junior high Negroes appeared to be least concerned with anxiety and frustration in meeting job qualifications. Significant differences, at the .05 level, were found with junior high Caucasians being more concerned over portrayed figures lacking training, skill, and education than senior high Negroes. Job Corps Negroes were significantly more concerned

over combined training, skill, and education and lack of training, skill and education, .05 level, than junior high Negroes.

Junior high Caucasians were significantly more concerned over the future, .05 level, than senior high Negroes and senior high Caucasians. There were no significant differences between groups for concern over possible denial of prestige.

Senior high Negroes were significantly more concerned over possible denial of achievement than Job Corps Negroes, .01 level; Job Corps Caucasians, .01 level; junior high Negroes, .02 level; and junior high Caucasians, .05 level. Senior high Caucasians were significantly more concerned than Job Corps Negroes, .01 level; junior high Caucasians, .01 level; and junior high Negroes, .02 level. Total senior high pupils were significantly more concerned with possible denial of achievement than total junior high pupils, .001 level.

There were no significant differences between groups in proportions of subjects falling into the delayed as compared to the immediate need gratification categories. The results suggested that concern for security acts as a major deterrent to delayed need gratification.

The subjects were asked to rank the stimulus cards in the order they liked best. It was found that the Blank Card was ranked first. The results suggested that the responses to the Blank Card were probably a reliable indicator of the subjects' occupational identification level. The wide variety of responses to the cards, though descriptive in nature, indicated that the subjects were responding to their individual perceptions and needs.

The results suggested that age and educational level were contributing factors in the differences found between group means for response time, cue time, and total response time. Significant differences at the .01 level were found between the following groups: Junior high Caucasians used less time than Job Corps Negroes, junior high Caucasians used less time than senior high Caucasians, and total junior high pupils used less time than total Job Corps enrollees.

Cue time was defined as the time between the presentation of the stimulus card and the subject's first response to the card. Significant differences at the .001 level were found between the following groups: Junior high Negroes required more time to respond

than Job Corps Negroes and Job Corps Caucasians; junior high Caucasians required more time to respond than Job Corps Caucasians; junior high Negroes required more time than senior high Negroes and senior high Caucasians; total junior high pupils required more time to respond than total Job Corps enrollees; and total junior high pupils required more time to respond than total senior high pupils.

Total response time was the length of time taken to complete the total test: a combination of response time and cue time. Significant differences at the .01 level were found between the following groups, with Job Corps Caucasians requiring more time than junior high Negroes and junior high Caucasians, and with total Job Corps enrollees requiring more time than total junior high pupils. Significant differences at the .05 level were found with Job Corps Negroes requiring more time than junior high Negroes, and with senior high Negroes requiring more time than junior high Caucasians. Differences between groups on total response time appeared to be related to age and work experience.

The results of this study must be viewed as tentative in nature due to the exploratory aspects of the POAT, a lack of data from other groups for purposes of comparison, the small size of the sample, and a lack of similar instruments which could be used for purposes of validation. However, the POAT seemed to provide a method of studying and analyzing the attitudes and perceptions of culturally disadvantaged youth towards work.

REFERENCES

- Allport, G. W. Attitudes in C. Murchison (ed.) *A handbook of social psychology*. Worcester: Clark University Press, 1935.
- Allport, G. W. *The nature of prejudice*. Cambridge: Addison-Wesley Press, 1954.
- Allport, G. W. *Personality and social encounter*. Boston: Beacon Press, 1964, 30.
- Ammons, R. B., Butler, M. N., and Herzog, S. A. *The vocational apperception test*. Louisville: Southern University Press, 1951.
- Anderson, H. H. and Anderson, G. L. *An introduction to projective techniques*. New York: Prentice-Hall, 1951.
- Bloom, B. S., Davis, A., and Hess, R. *Compensatory education for cultural deprivation*. New York: Holt, Rinehart and Winston, 1965.
- Forer, B. R., A diagnostic interest blank. *Journal of Projective Techniques*, 1948, 12, 119-129.

- Forer, B. R. Framework for the use of clinical techniques in vocational counseling. *Personnel and Guidance Journal*, 1965, 43, 868-872.
- Frank, G. H. Biochemical implications. *American Psychologist*, 1964, 19, 54.
- Goodstein, L. D. Some comments on clinical appraisal in vocational counseling. *Personnel and Guidance Journal*, 1965, 43, 882-883.
- Lazarus, R. S. *Personality and adjustment*. Englewood Cliffs: Prentice-Hall, 1963, 60-63.
- McCabe, S. P. Hazards in the use of clinical techniques in vocational counseling. *Personnel and Guidance Journal*, 1965, 43, 879-881.
- McCloskey, G. and Olsen, L. C. *Research and development: Proposal for a second stage of vocational-technical education, research, and development project*. ERD-257-65, Washington, U.S. Department of Health, Education, and Welfare, Jan., 1966.
- Miller, S. M. and Riessman, F. Social class and projective tests. *Journal of Projective Techniques*, 1958, 22, 443-439.
- Mindess, H. Psychological indices in the selection of student nurses. *Journal of Projective Techniques*, 1957, 21, 37-39.
- Office of Economic Opportunity, *Job Corps Screening Handbook*, Washington, D.C. Office of Economic Opportunity, 2, 6, 24-26.
- Office of Strategic Services Assessment Staff, *Assessment of men*. New York: Rinehart, 1948.
- Olsen, L. C. *Development and standardization of a projective occupational attitude test*, Research Report. Project No. ERD-257-65, Contract No. OE-5-85-109, Washington, U.S. Dept. of Health, Education, and Welfare, Office of Education, Bureau of Research, December, 1966, 5, 18.
- Olsen, L. C. *Development of a projective technique for obtaining educationally useful information indicating youths' attitudes toward work and occupational plans*, Research Report. Project No. OE7-0031, Contract No. OEG-4-7-070031-1626, Washington, U.S. Dept. of Health, Education, and Welfare, Office of Education, Bureau of Research, June, 1968, 26-29.
- Phelan, J. G. Projective techniques in the selection of management personnel. *Journal of Projective Techniques*, 1962, 26, 102-104.
- Roe, A. A new classification of occupations. *Journal of Counseling Psychology*, Winter, 1954, p. 215-220.
- Roe, A. Personality structure and occupational behavior. *Man and the world of work*. Editor H. Borrow, Boston: Houghton Mifflin, 1964.
- Roe, A. and Mierzwa, J. The use of the Rorschach in the study of personality and occupations. *Journal of Projective Techniques*, 1960, 24, 282-289.
- Steiner, M. E. The search for occupational personalities: The Rorschach test in industry. *Personnel*, 1953, 29, 335-343.

THE VALIDITY OF MEASURES OF EYE-CONTACT

MARVIN E. SHAW, J. THOMAS BOWMAN, AND
FRANCES M. HAEMMERLIE

University of Florida

THE significance of mutual eye-contact for interpersonal behavior has been noted by a number of writers (e.g., Simmel, 1921; Heider, 1958), but it was not until Exline's (1963) study that this phenomenon began to be examined under controlled conditions. Since that time, numerous experimental studies have been reported concerning the relationship of eye-contact to other aspects of the interpersonal situation (Exline, Gray, and Schuette, 1965; Argyle and Dean, 1965; Efran, 1968; Efran and Broughton, 1966). In the typical experiment, two persons engage in conversation during which the frequency and duration of eye-contact is recorded by observers located behind a one-way vision screen. Observers record eye-contact by means of push-buttons which operate pens on an event recorder. It is, therefore, possible to compare directly the scores of two or more observers, thus obtaining evidence of the reliability of such measures. The evidence suggests that good inter-observer reliability can be achieved; for example, Exline (1963) reported inter-observer reliabilities of .97 to .98.

Curiously enough, little attention has been given to the question of the validity of such measures; i.e., the degree to which such measures reflect actual eye-contact. Exline (1963) cited an unpublished study by Gibson and Davidson (later published by Gibson and Pick, 1963) as evidence that two individuals can judge accurately when they are looking into each other's eyes. This same evidence was cited by Argyle and Dean (1965). Unfortunately, examination of the report by Gibson and Pick (1963) reveals that this study did

not deal with eye-contact at all. Instead, Gibson and Pick were interested in line of gaze; they demonstrated that subjects can judge accurately whether another person is fixating above, below, or on the bridge of the nose.

There appears to be only one other study that referred to the validity of such observations. Exline, Gray, and Schuette (1965) reported a pretest in which an interviewer maintained constant fixation on the subject's eyes, while both he (the interviewer) and an observer recorded frequency and duration of the subject's visual fixations on the interviewer. Since the interviewer was always available for eye-contact, one may assume that mutual eye-contact occurred each time the subject fixated on the interviewer's eyes. Exline et al. reported that the observer and interviewer agreed on 88 per cent to 98 per cent of their judgments. They gave no details concerning distance between interviewer and subject, distance between observer and subject, number of subjects, or how agreement was computed.

The purpose of the present study was to provide empirical evidence concerning some of the implicit assumptions about eye-contact and its measurement that are inherent in the kinds of experiments cited above, namely, (a) that two individuals can judge accurately the frequency and duration of mutual eye-contact, and (b) that an observer can validly judge when two other persons are looking into each other's eyes. Since distance between the members of the dyad has been shown to be an important determinant of eye-contact as usually measured (Argyle and Dean, 1965), data were collected at several inter-person distances.

Study 1

Method

The first part of this investigation attempted to determine whether two persons could judge accurately when they were looking into each other's eyes. Two persons, one male and one female, who had previously served as observers in an eye-contact study, engaged in two ten-minute discussions at each of four inter-person distances: 2½ ft., 4½ ft., 8 ft., and 12 ft. Each person had a concealed push-button which was connected to an Esterline Angus Event Recorder located in an adjacent room. During the discussion, each person

pushed his button each time he believed that he had made eye-contact with the other person and released it when he believed eye-contact was broken. Since neither knew when the other had pushed his button, each person's observation was independent of the other person's observation, except that both depended upon mutual eye-contact. Agreement between these two judgments presumably reflects actual eye-contact.

Results and Discussion

Each discussion period was divided into one-minute units and frequency and duration scores computed for each unit. The correlations between the two sets of frequency scores (number of eye-contacts per minute) in each session are shown in Table 1 and correlations between duration scores (amount of eye-contact per minute) are given in Table 2. As can be seen, correlations ranged from .59 to 1.00, with average correlations (computed by z transformation procedures) ranging from .85 to .98. Agreement regarding duration of eye-contact was significantly less at $2\frac{1}{2}$ ft. than at 8 ft. ($p < .03$) or at 12 ft. ($p < .05$). This effect was due entirely to the low correlation obtained in the first session. This low agree-

TABLE 1

Correlations between Frequency of Eye-Contact Reported by Members of a Dyad as a Function of Distance

	Distance between Members of the Dyad			
	$2\frac{1}{2}$ ft.	$4\frac{1}{2}$ ft.	8 ft.	12 ft.
Session 1	.78	.97	1.00	.92
Session 2	.96	.80	.87	.92
Mean	.91	.92	.97	.92

TABLE 2

Correlations between Duration of Eye-Contact Reported by Members of a Dyad as a Function of Distance

	Distance between Members of the Dyad			
	$2\frac{1}{2}$ ft.	$4\frac{1}{2}$ ft.	8 ft.	12 ft.
Session 1	.59	.91	.98	.98
Session 2	.95	.94	.95	.91
Mean	.85	.93	.98	.96

ment could have been caused by lack of adjustment to the recording procedure, or to the fact that such close proximity was relatively uncomfortable—a fact noted by the participants.

In general, however, the correlations between judgments of participants were as high as could reasonably be expected, assuming some error of judgment due to momentary distraction or involvement in the discussion topic. Hence, we believe that the evidence strongly supports the common belief that two persons know when they are looking into each other's eyes.

Study 2

Method

The second part of this investigation was designed to determine the degree of correspondence between judgments of mutual eye-contact made by a participant and judgments made by an observer via a one-way vision screen. High correlations would provide strong evidence that scores obtained by means of observers do indeed validly measure mutual eye-contact. Since such measures are often obtained at varying observer-participant distances, data were collected at three such distances: 4½ ft., 8 ft., and 12 ft. The 2½ ft. distance was omitted because of the impracticality of trying to get the participant-subject that close to an observer located on the opposite side of the one-way vision screen.

In this study, the two persons who had participated in Study 1 served as interviewer and observer. Each person played each role an equal number of times. As interviewer, each person was paired with a naive member of the opposite sex. The interview situation was arranged so that the interviewer sat with his back to the one-way screen; the subject faced both the screen and the observer located behind it. The interviewer engaged the subject in a discussion of any topic of mutual interest for a ten-minute period, during which time both the interviewer and the observer recorded eye-contact via push-buttons and the event recorder. The interviewer attempted to make eye-contact available to the subject at all times, without appearing to be staring at him. Six interviews were conducted at each of the three subject-observer distances, for a total of 18 interviews. The distance between the interviewer and the subject was approximately 4½ ft. under all conditions.

Results and Discussion

Each interview was again divided into one-minute units, and frequency and duration scores computed for each unit. Correlations were computed between the scores of the interviewer and those of the observer. Table 3 gives the correlations for frequency scores and Table 4 presents the correlations for duration of eye-contact.

Correlations were noticeably lower here than in the first study, ranging from .55 to .98 for frequency scores and from .56 to .98 for duration scores. Variations among sessions at the same distance are substantial, and the average correlations are noticeably smaller at the greater observer-subject distances. Correlations between participant and observer duration scores were significantly higher ($p < .05$) at 4½ ft. than at either 8 ft. or 12 ft.; correlations for frequency scores were significantly greater at 4½ ft. than at 12 ft.

TABLE 3

Correlation between Frequency of Eye-Contact Recorded by an Observer and Reported by a Member of the Dyad

	Distance between Subject and Observer		
	4½ ft.	8 ft.	12 ft.
Male	.97	.78	.82
Subjects	.85	.87	.85
	.64	.88	.55
Female	.66	.58	.95
Subjects	.98	.93	.81
	.92	.79	.80
Mean	.90	.83	.82

TABLE 4

Correlation between Duration of Eye-Contact Recorded by an Observer and that Reported by a Member of the Dyad

	Distance between Subject and Observer		
	4½ ft.	8 ft.	12 ft.
Male	.92	.68	.74
Subjects	.96	.97	.91
	.70	.82	.83
Female	.98	.67	.56
Subjects	.75	.61	.96
	.95	.94	.88
Mean	.92	.84	.85

($p < .05$) and differences between correlations at $4\frac{1}{2}$ ft. and 8 ft. approached significance ($p < .08$).

One plausible hypothesis to account for the variability from session to session might be that validity of judgments increased with practice; however, the rank order correlation between session-order and observer-interviewer correlations was not significant for either duration or frequency scores. It might also be suspected that the subject sometimes looked into the eyes of the interviewer when the interviewer was looking elsewhere. If so, the interviewer would not record a contact whereas the observer would. This could account for both variability and the relatively low interviewer-observer correlations. Two bits of evidence mitigate against this interpretation. First, the interviewer attempted to provide constant eye-contact availability. Second, the mean frequency and duration scores of the interviewer were almost identical to those of the observer, whereas the above interpretation requires that mean observer scores be higher than mean interviewer scores.

Therefore, the variability appears to be due to uncontrolled variations in the situation from session to session. Since the procedures followed were made as uniform as possible, it seems probable that similar uncontrolled variations occur in other studies of eye-contact.

The easy explanation of decreased validity of judgments with increased observer-subject distance is that judgments become more difficult with distance. Unfortunately, there was no good method for testing this hypothesis.

Implications for Research

There are three main conclusions that can be drawn from this research:

1. An individual can judge accurately when he is looking into another's eyes. This conclusion is inferred from the finding that members of a dyad show high agreement concerning mutual eye-contact.
2. The validity of observer's judgments of eye-contact vary considerably from session to session, where validity is inferred from agreement between observer and participant.
3. The validity of observer's judgments of eye-contact varies with the distance between the observer and the subject.

The first of these merely verifies a common belief about an individual's ability to know when he is making eye-contact with another person, and provides a basis for evaluating the validity of the observer's judgments of eye-contact. The second two, however, have important implications for most research using eye-contact as a dependent variable. Since the validity of observer's judgments varies from time to time, it is incumbent upon the investigator to demonstrate that the particular measures that he uses are reliable and valid. Even if this variability is limited to the two particular observers that were used in this study, this means that not all trained observers are equally accurate in their judgments of eye-contact. Hence, each investigator must demonstrate that his observer(s) can make valid judgments under the conditions of his investigation. It is not sufficient to cite the results of previous studies using different observers.

The consequences of differential validity at different observer-subject distances apply primarily to studies using distance as an independent variable. The fact that measures of eye-contact are not equally valid for all such distances must be taken into account in interpreting the findings of such investigations.

REFERENCES

- Argyle, M. and Dean, J. Eye-contact, distance, and affiliation. *Sociometry*, 1965, 28, 289-304.
- Efran, J. S. Looking for approval: Effects on visual behavior of approbation from persons differing in importance. *Journal of Personality and Social Psychology*, 1968, 10, 21-25.
- Efran, J. S. and Broughton, A. Effects of expectancies for social approval on visual behavior. *Journal of Personality and Social Psychology*, 1966, 4, 103-107.
- Exline, R. V. Explorations in the process of person perception: Visual interaction in relation to competition, sex, and the need for affiliation. *Journal of Personality*, 1963, 31, 1-20.
- Exline, R. V., Gray, D., and Schuette, D. Visual behavior in a dyad as affected by interview content and sex of respondent. *Journal of Personality and Social Psychology*, 1965, 1, 201-209.
- Gibson, J. J. and Pick, A. D. Perception of another person's looking behavior. *American Journal of Psychology*, 1963, 76, 386-394.
- Heider, F. *The Psychology of interpersonal relations*. New York: Wiley, 1958.
- Simmel, G. Sociology of the senses: Visual interaction. In R. E. Park & E. W. Burgess (Eds.), *Introduction to the science of sociology*. Chicago: University of Chicago Press, 1921.

VALIDITY STUDIES SECTION

WILLIAM B. MICHAEL, Editor
University of Southern California

JOAN J. MICHAEL, Assistant Editor
California State College, Long Beach

<i>Combining the Ipsative and Normative Approaches in Selection Validation.</i> MARGARET A. HOWELL	931
<i>Validity and Likability Ratings for Three Scoring Instructions for a Multiple-Choice Vocabulary Test.</i> CARRIE WHERRY WATERS AND LAWRENCE K. WATERS	935
<i>Advanced Placement Scores: Their Predictive Validity.</i> PAUL S. BURNHAM AND BENJAMIN A. HEWITT	939
<i>SAT and High School Average Predictions of Four Year College Achievement.</i> MARVIN SIEGELMAN	947
<i>The Relationship Between Expected Grades and Students' Evaluations of Their Instructors.</i> DAVID S. HOLMES	951
<i>Factor Analysis of 1970-71 Version of the Comparative Guidance and Placement Battery.</i> JOSEPH GRIMALDI, EUGENE LOVELESS, JAMES HENNESSY AND JOHN PRIOR	959
<i>Correlates of a Pass-Fail Decision for Admission to Candidacy in a Doctoral Program in Education.</i> WILLIAM B. MICHAEL, ROBERT A. JONES, HUDHAIL AL-AMIR, CALVIN M. PULLIAS, MICHEL JACKSON, AND VALERIE GOO	965
<i>Appropriateness of Subtests in Achievement Tests Selection.</i> THOMAS M. GOOLSBY, JR.	969
<i>The Reliability and Validity of the Contemporary Mathematics Test.</i> JOSEPH S. RENZULLI AND ROBERT A. SHAW ..	973
<i>Differential Effects of Initial Course Placement as a Function of ACT Mathematics Scores and High School Rank-in-Class in Predicting General Performance in Chemistry.</i> JOHN R. REINER	977
<i>The Criterion-Related Validities of Cognitive and Non-Cognitive Predictors in a Training Program for Nursing Candidates.</i> WILLIAM B. MICHAEL, RUSSELL HANEY, YOUNG B. LEE, AND JOAN J. MICHAEL	983

<i>A Note on the Validity of Two Measures of High School Rank.</i>	
GERALD W. McLAUGHLIN	989
<i>Multivariate Validity of the Otis-Lennon Mental Ability Tests Primary I Level.</i>	
BRAD S. CHISSOM AND JERRY R. THOMAS	991
<i>The Concurrent Validity of the Sprigle School Screening Readiness Test for a Sample of Preschool and Kindergarten Children.</i>	
MARIA S. A. SEDA AND JOAN J. MICHAEL	995
<i>A Multitrait-Multimethod Validation of Measures of Student Attitudes Toward School, Toward Learning, and Toward Technology in Sixth Grade Children.</i>	
SOL M. ROSHAL, IRENE FRIEZE, AND JANET T. WOOD	999
<i>Dogmatism and Conservatism: An Empirical Follow-up of Rokeach's Findings.</i>	
FRANK COSTIN	1007
<i>Specific Anxiety Theory and the Mandler-Sarason Test Anxiety Questionnaire.</i>	
FRANK B. W. HARPER	1011
<i>Hostility and Learning: A Follow-up Note.</i>	
FRANK COSTIN ..	1015
<i>Predicting the Behavior of Institutionalized Delinquents With—and Without—Cattell's HSPQ.</i>	
VERNON O. TYLER, JR. AND ROBERT F. KELLY	1019
<i>Interest Profiles of Clergymen as Indicated by the Vocational Preference Inventory.</i>	
DAVID L. SCHULDT AND ROBERT F. STAHMANN	1025

PROVISION FOR PUBLICATION OF VALIDITY STUDIES OF ACADEMIC ACHIEVEMENT

This section is provided for the early publication, at the expense of the author, of short validity studies of academic achievement. The charge to the author is forty-five dollars per page plus ten dollars extra per page for tables, figures, and formulas. Authors are granted permission to have reprints made of their articles at their own expense.

Preference will be shown for manuscripts of fewer than 1200 words, with no more than six references and containing two or fewer tables each of no more than one 8½" x 11" elite typed page—making six printed pages. Any manuscript exceeding 3000 words, 12 references, and four tables or figures equivalent to three 8½" x 11" typed pages will be automatically returned, as 12 printed pages will be the maximum total number of pages for any article to be published in this section.

The Validity Studies of Academic Achievement Section is published twice a year, once in the Summer issue and again in the Winter issue, for which the closing dates for receiving manuscripts are December first and June first, respectively.

Two copies of the manuscripts should be sent to:

Dr. William B. Michael
325 Callita Place
San Marino, California 91108.

COMBINING THE IPSATIVE AND NORMATIVE APPROACHES IN SELECTION VALIDATION

MARGARET A. HOWELL

Department of Health, Education, and Welfare

THE traditional model of selection validation is that of a predictor validated against a criterion, a measure of job performance. Elaborations of the model include multiple correlation involving more than a single predictor and canonical correlation which deals with multiple criteria as well as multiple predictors.

Basic to the classical validation model is the concept of reliability of measurement in which it has been assumed that an obtained score is composed of the true score plus random error variance (Guilford, 1954, p. 349). Increasing evidence, such as research on moderator variables, suggests that instead of random, the error variance is systematic and errors of prediction are predictable (Ghiselli, 1963). A moderator, though based on recognition of nonrandom error variance, sorts individuals into subgroups to which the classical validation model is then applied. Subgroup sorting involves assigning individuals to "types" for selection purposes.

Rather than the use of a moderator variable based on normative measurement, the ipsative approach has become associated with typologies (Block, 1961, p. 16). Ipsative measurement is allied with the idiographic rather than the nomothetic view of psychology (Allport, 1961, p. 8). The *Q* sort, based on descriptive statements, is like an ipsative item of the forced-choice kind in that the focus is on the way in which traits are organized intra-individually. The *Q* technique, involving correlations among individuals, is the correlational counterpart of the ipsative item. In the application of the *Q* technique, the correlations among pairs of individuals represent a re-standardizing of the standard scores derived from the normative

approach to reflect score deviations around individual means of zero. Factor-analysis of a Q correlational matrix allows the identification of types in that individuals of the same factor type display a similar intra-individual pattern of traits even though the level of the traits, in a normative sense, may differ.

One author has warned against confounding ipsative and normative measurement (Broverman, 1962, p. 295). It could be argued, however, that the only meaningful intra-individual variability, other than a mere ordering of traits, must first be expressed in terms of normative measurement. Even with ipsative items, this would suggest the items might be scaled normatively prior to their ipsative use.

Proposed combined model—For selection purposes, perhaps the "saliency" of traits within the individual as well as inter-individual variability needs recognition. This would mean that the classical selection model might be replaced by one combining the ipsative and normative approaches. The concern in validation would be predicting either for the individual or for the type and not "on the average" as is represented by the traditional validity coefficient. Instead of a regression model in which the same weights are applied to all individuals, weights would vary by typology. Q analysis of multiple criteria affords a means of arriving at the subgroups or types for which different weights or even different predictor variables might be used for validation purposes. Validity could be determined in terms of profile similarity between N predictor and N criterion measures for an individual, although there are technical difficulties in the use of measures of profile similarity (Guion, 1965, p. 174). For each subgroup, an average index of profile similarity between predictor and criterion profiles could be obtained.

If the combined ipsative and normative selection model tested out adequately in both concurrent and predictive validation studies, its use in practice would be similar to administering a selection-placement battery. On a new group of applicants, scores on all predictors would be obtained. Various profiles relevant to the identified criterion types would be developed on each applicant. The profile yielding the best match with a particular subgroup would indicate appropriate placement for the individual. If, for example, Q analysis of criterion measures identified a subgroup of low speed and high accuracy typists who were also identifiable from a subset of

predictor variables, an applicant fitting this type could be placed in a position requiring this intra-individual organization of traits.

Implications of the model—This model for selection validation assumes individuals are unique, genetically and behaviorally, but can be subgrouped into typologies, where a type has meaning only to the extent it is useful for measurement purposes. With sufficiently refined and comprehensive measurement, the subgroup N would equal 1!

REFERENCES

- Allport, G. W. *Pattern and growth in personality*. New York: Holt, Rinehart, and Winston, 1961.
- Block, J. *The Q-Sort method in personality assessment and psychiatric research*. Springfield, Ill.: Charles C Thomas, 1961.
- Broverman, D. M. Normative and ipsative measurement in psychology. *Psychological Review*, 1962, 69, 295-305.
- Ghiselli, E. E. Moderating effects and differential reliability and validity. *Journal of Applied Psychology*, 1963, 47, 81-86.
- Guilford, J. P. *Psychometric methods*. (2nd ed.). New York: McGraw-Hill, 1954.
- Guion, R. M. *Personnel testing*. New York: McGraw-Hill, 1965.

VALIDITY AND LIKABILITY RATINGS FOR THREE SCORING INSTRUCTIONS FOR A MULTIPLE-CHOICE VOCABULARY TEST

CARRIE WHERRY WATERS

Center for Psychological Services

Ohio University

LAWRENCE K. WATERS

Ohio University

TYPICAL instructions for discouraging examinees from guessing on multiple-choice tests indicate that some fractional amount will be subtracted for each wrong answer. The intent is to encourage examinees to omit items they do not know. Recently, Traub, Hambleton, and Singh (1969) suggested the addition of some fractional number of points for omitted items as a more direct and effective method of eliciting the desired test-taking behavior. They presented evidence of higher inter-form reliability, and an increase in the number of omitted items, for the "reward for omits" instructions as compared to the "penalty for wrongs" instructions.

It also seems reasonable that more examinees would prefer the direct positive reinforcement of omissive behavior approach to the less direct negative reinforcement approach. The purpose of the present study was to compare these two approaches and a third "rights only" scoring method in terms of (a) how well each method is liked by examinees and (b) the correlations of scores obtained under each method with test and course performance criteria.

Method—Subjects. A total of 72 students (48 females and 24 males) from two educational psychology classes participated in the study during a regular class period. The subjects were divided into three groups of 24 examinees each for the purposes of the study.

Tests. Three matched vocabulary tests of 15 items each were con-

constructed from an item pool for which p -values were available. The corresponding items in the three tests were matched on difficulty to within $\pm .05$. The item difficulty distributions were skewed with more items in the higher difficulty range. For each item, the examinee was required to choose one of the five alternative words which was most nearly opposite in meaning to the stem word.

Instructions. All examinees were given the same instructions for answering the items. In addition, three sets of instructions concerning scoring of the tests were given. The scoring procedures (in terms of the weights for rights, wrongs, and omits respectively) were 1, 0, 0; 1, $-\frac{1}{4}$, 0; and 1, 0, $\frac{1}{5}$. These scoring procedures were chosen to represent the no penalty for guessing condition, the conventional correction procedure for random guessing on five alternative items, and the procedure suggested by Traub et al. (1969). Each of the scoring procedures was presented to the examinee in a table showing the number of points gained or lost for a right, wrong, or omitted item. Following the table, a further explanation of the particular scoring procedure was given (e.g. "That is, for each correct answer you get one point, for each wrong answer you lose $\frac{1}{4}$ of a point, and an omitted—unanswered—item does not count for or against you"). The instructions for each test were given on a separate page immediately preceding the test items.

Test booklets. In order for all examinees to respond to items under each of the three instructions, a 45-item "Verbal Ability Test" booklet was constructed. Each section of the booklet consisted of one of the three instructions paired with one of the three 15-item tests. A 3×3 Latin Square was used to construct the booklets (rows = groups, columns = instructions, and cells = 15 item tests). Although not analyzed, the three instruction-test combinations for each group were given in all six possible orders to an equal number of examinees.

The last page in each booklet was an evaluation sheet. Examinees first ranked (1—most liked, 3—least liked), then rated on a 9-point scale (9—liked very much, 1—disliked very much), how well they liked each section of the test booklet.

The 72 test booklets were shuffled before being distributed to the students. No time limit was imposed and all examinees finished within the class period.

Results and discussion—Liking rankings and ratings. The mean

ranks for the rights only, $\frac{1}{5}$ of a point for omits, and $-\frac{1}{4}$ of a point for wrongs instructions were 1.72, 1.82, and 2.46 respectively. A Friedman χ^2 computed on the ranked data across all 72 examinees was 23.74 ($p < .001$, $df = 2$). Thus, in terms of relative preference, the $-\frac{1}{4}$ for wrongs instructions were consistently ranked lower than either of the other instructions.

A summary of the analysis of variance of the ratings is given in the left-hand columns of Table 1. The only significant F was for Instructions.

Using the Newman-Keuls procedure, comparisons among the instructions' means indicated that the rights only (4.89) and $\frac{1}{5}$ for omits (4.67) instructions were rated significantly higher than the $-\frac{1}{4}$ for wrongs instructions (3.36). These data support the use of a procedure of allowing some fractional number of points for omitted items, rather than the conventional procedure for discouraging guessing on multiple-choice tests.

Number of omitted items—An analysis of variance of the number of omitted items (summarized in the right-hand columns of Table 1) indicated that both Groups and Instructions were significant. No explanation can be offered for the difference among the groups (Group 3 omitted significantly more items than Groups 1 and 2). A comparison of the means for the three instructions using the Newman-Keuls procedure showed that the $\frac{1}{5}$ for omits (6.29), and $-\frac{1}{4}$ for wrongs (6.42) instructions resulted in significantly more omitted items than the rights only instructions (3.00).

Because the number of omits for the rights only instructions

TABLE 1
Summary of Analysis of Variance for Ratings of Liking and for Number of Omitted Items

Source	df	RATINGS		OMITS	
		MS	F	MS	F
Between Examinees	71				
Groups (G)	2	10.17	1.27	96.02	6.19*
Error _b	69	7.98		15.52	
Within Examinees	144				
Instructions (I)	2	49.06	12.84*	270.29	26.29*
Tests (T)	2	3.94	1.03	22.26	2.17
(IxT)	2	1.72	<1	5.60	<1
Error _w	138	3.82		10.28	

* $p < .01$.

seemed somewhat high, a comparison was made of the number of omits under these instructions when the instructions were for the first, second, and third sections of the booklets. When the rights only instructions were for the second or third sections, almost twice as many items were omitted as when the instructions were for the first section of the booklet. It is hypothesized that an omissive set, from either or both of the instructions designed to elicit omits, carried over to the rights only section when it followed either or both of the other instructions. This effect was not found for the other two instructions.

Correlations with test and course performance criteria—To compute the correlations involving scores obtained under each of the instructions, the three matched 15-item tests were considered as equivalent forms. Three right scores (one for each instruction) and two formula scores ($R + \frac{1}{5} O$ and $R - \frac{1}{4} W$) were computed for each examinee. Since the formula scores correlated .95 with the rights scores obtained under corresponding instructions, only data for the rights scores are reported. The scores obtained under "rights only," " $\frac{1}{5}$ of a point for omits," and " $-\frac{1}{4}$ of a point for wrongs" instructions correlated .63, .50, and .59 with SAT-V ($n = 40$) and .42, .30, and .29 respectively with course grade ($n = 70$). None of the correlations for the three instructions differed significantly from each other ($p > .05$).

In general, more items were omitted under both the bonus for omits and penalty for wrongs instructions than under the rights only instructions and no significant differences were found in the correlations of scores obtained under the instructions with test and course performance criteria. However, the sections of the test taken under the bonus for omits instructions were liked better than those taken under the penalty for wrongs instructions. If guessing behavior is to be discouraged on a multiple-choice test, the results of Traub, et al. (1969) and the present study seem to indicate that a procedure for rewarding the examinee for omitted items would be superior to penalizing the examinee for wrong answers.

REFERENCE

- Traub, R. E., Hambleton, R. K., and Singh, B. Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 847-861.

ADVANCED PLACEMENT SCORES: THEIR PREDICTIVE VALIDITY¹

PAUL S. BURNHAM AND BENJAMIN A. HEWITT

Yale University

WHILE Advanced Placement scores have been used increasingly since the inception of the program in 1952, little "hard data" evidence of their validity has been published. Developed by the College Entrance Examination Board, the program functions under the assumption that able students can successfully complete some college courses while they are in secondary school (Blackmer, 1952; Casserly, 1966; Cornog, 1956; and Wilcox, 1962). Participating schools offer to their better students courses which the program has planned, outlined, and described. These students take examinations which the program's committee of readers grades on a scale ranging from five, extremely well qualified, to one, no recommendation. After participating colleges receive the scores they may decide to place students in advanced courses or to award them college credit.

To judge the value of this program we decided to relate the AP scores of Yale students to criteria of ability and achievement, such as scores on other tests and grades in pertinent courses and programs. We were disappointed to find that extensive use of AP scores in course placement had resulted in the severe fractionization of many groups which had appeared initially sizeable and promising. For example, the class of 1963, with which we started to work, was comprised of 1031 matriculants; they had submitted 533 AP scores. Of these, only the 149 in English and the 128 in

¹ This study was partially supported by a research grant from the College Entrance Examination Board.

mathematics could be analyzed profitably. Preliminary work revealed that regardless of the score they achieved, the students taking the AP examination in English and/or in mathematics were initially more able and produced better as Yale freshmen than the average matriculant.

To obtain more evidence we accumulated the same kinds of information for the class of 1967, where we had a study group of 1027 students; 500 had presented 930 AP scores in the seven examination fields appearing sizeable enough to study.

Ability of scores to differentiate performance level—In courses in the field of the AP examination the average grades of AP students rose as their AP scores increased. Using class of 1967 data, we accumulated all of the freshman and sophomore course grades of the AP students in each field except physics; the latter numbered too few students to warrant this kind of analysis. Next these grades were averaged after they had been grouped according to the field of the man's examination and his score.

Rankings by magnitude of the AP scores and of the average college grades in that field corresponded highly with only minor exceptions. These findings were particularly impressive because the data included grades of many high-scoring students who took advanced courses initially and even higher-level courses later; included too, were grades of students who took only the first course offered by a department. Thus, these mean grades reflected the performance of students in courses at various levels of advancement and complexity and with disparate content and grading standards.

Multiple correlation evidence of the predictive power of scores—Of all the courses in Yale College, only the six shown in Table 1 had enough students with AP scores of one through five to justify a multiple correlation analysis involving the student's term-course grade as the dependent variable and his SAT-V score, his predicted freshman-year average, and his scores on relevant CEEB achievement and aptitude tests as the independent or predictor variables. The latter are indicated in Table 1. By adding the AP score to the other independent variables, we found a slight increase in each of the six multiple correlations. While the samples are small, the consistency with which the AP scores made a useful

TABLE 1
Multiple Correlation Data, Class of 1967

Course ^a	Predictor Variables in Addition to SAT-V and Matriculation Prediction ^b	Multiple Correlation with Grades in Yale Courses	
		Excluding AP Scores	Including AP Scores
English 24 (<i>n</i> = 35)	CEEB English	.18	.22
English 25 (<i>n</i> = 87)	AP English		
	CEEB English	.34	.38
	AP English		
French 41 (<i>n</i> = 17)	CEEB French	.46	.49
	AP French		
Mathematics 10 } (<i>n</i> = 21)	CEEB SAT-M,	.64	.73
Mathematics 15 } (<i>n</i> = 71)	Adv. Math.	.41	.50
Mathematics 20 } (<i>n</i> = 30)	AP Math	.39	.42
	Median Correlation	.40	.45

^a Courses in Chemistry, History and Physics were not analyzed because of the paucity of data.

^b A prediction of each student's probable average grade in all his Freshman year courses.

contribution to the other matriculation data offers further evidence of their predictive validity.

Ability and achievement of AP and non-AP students. In each subject except French Literature, students making AP scores of four and five were superior to other AP students in ability as measured by the mean of each student's CEEB scores. In turn, in all seven AP subject-matter fields, the group with AP scores proved to be superior in their CEEB records to students who had attended "AP schools" but did not take that particular AP test.

AP students attained higher average freshman and sophomore grades within the field of the AP test than students who attended "AP schools" but did not take the test. As one might expect, the high-scoring AP men made the more impressive records in the field of the examination; only in physics was this not the case. Table 2 provides supporting data.

Matched-group comparisons—The extent to which college-level work can be completed effectively in secondary school might be inferred by comparing the grades of AP freshmen assigned to an advanced course with those of non-AP sophomores having comparable ability in terms of CEEB averages, enrolled in the same

TABLE 2

*Summary of Matriculation and College Performance Records of Class of 1967 Students
AP Secondary Schools*

Advanced Placement Subject	Offered AP Scores				Did Not Offer AP Score
	AP 4, 5		AP 1-3		
<i>English</i>					
Number of Matriculants	64		228		240
CEEB Average ^a	690		664		625
Average English Grade ^b	83.6		81.3		78.1
<i>Chemistry</i>					
Number of Matriculants	35		51		159
CEEB Average ^a	715		673		650
Average Chemistry Grade ^b	79.9		78.6		75.0
<i>French</i>					
	<i>La.</i>	<i>Lit.</i>	<i>La.</i>	<i>Lit.</i>	
Number of Matriculants	26	12	28	42	194
CEEB Average ^a	675	665	656	665	630
Average French Grade ^b	83.1	83.1	79.8	80.9	77.4
<i>American History</i>					
Number of Matriculants	80		111		249
CEEB Average ^a	685		659		630
Average Amer. Hist. Grade ^b	84.1		81.7		77.3
<i>European History</i>					
Number of Matriculants	16		40		181
CEEB Average ^a	685		650		635
Average European Hist. Grade ^b	82.0		80.5		78.1
<i>Mathematics</i>					
Number of Matriculants	59		158		245
CEEB Average ^a	705		671		630
Average Mathematics Grade ^b	82.6		76.8		73.3
<i>Physics</i>					
Number of Matriculants	10		24		98
CEEB Average ^a	730		673		640
Average Physics Grade ^b	79.4		81.5		75.2

^a Based on both the Aptitude and Achievement Tests taken by each student.

^b Mean grade in all college courses taken in the field of the AP Test during freshman and sophomore years.

course, but entering it through a lower-level, two-term college sequence. (We matched students according to the mean of four or more of their CEEB scores; the SAT-V and -M comprised two of these, and the balance was formed by as many of the achievement test scores as each man had presented. Some of the means were based on as many as eight scores, but none on fewer than four. While we recognized that strict matching on the basis of the same four or more CEEB tests would have been more desirable, doing so would have fractionized our samples so much that this phase of the analysis could not have been accomplished.) We investi-

gated advanced courses to which AP students were initially assigned. Only in two English courses and one French course, each of two terms, and in one single-term mathematics course were we able to establish matched groups which could be compared by taking the average of their two term grades and their one grade in the single term course, as criteria of performance.

Data in Table 3 suggest that the freshmen who had gained advanced placement in English 25 and French 41 nearly "held their own" with the two groups of sophomores who had been prepared for these courses through a year of college study. In Mathematics 20, grades of the AP students were clearly superior to both groups of sophomores. This evidence, while limited to three subject-matter fields and involving only small samples, adds weight to other previously presented evidence that the AP Program can prepare able secondary-school students to compete effectively in upper-class college courses.

In English 15 we present information about students who made lower AP scores and were not assigned to upper-class courses; as a group they produced a bit better than the non-AP freshmen. One might expect them to do so because in a sense English 15 was a second exposure to the AP work they had done in secondary school but presumably was new material for the non-AP freshmen.

Discussion and conclusions. Numerous positive findings resulting from the study of small populations of Yale students offer consistent evidence of the validity of the AP Program.

1. As a group, AP students appeared more able and achieved better than the average matriculant. They were also superior to non-AP students from "AP schools."
2. In freshman and sophomore courses in the field of the AP examination the average grades of students grouped according to their AP score rose as the AP score increased and thus gave evidence of the predictive differentiation by the AP scores.
3. Since predictions of performance in college courses were improved slightly but consistently by adding AP scores to other CEEB tests of aptitude and achievement, AP tests appear to be measuring factors of some pertinence which are not otherwise included in the matriculation records.
4. Comparisons of the performance of AP freshmen with that of

TABLE 3

Achievement in Four College Courses of AP and Non-AP Students of Comparable Ability

	Class	N	Range of AP Scores	Average of Four or More CEEB Scores		Final Grade in Course	
				Mean	SD	Mean	SD
English 15, Problems in Writing and Literary Interpretation, the first college English course of:							
(a) Freshmen with English AP scores	1967	65	1-3	644	40	80.5	4
(b) Freshmen without English AP scores	1967	65		642	40	79.2	3
English 25, Major English Poets, Chaucer to Eliot							
(a) Freshmen with English AP scores taking this as first college English course	1967	16	2-5	618	31	80.8	5
(b) Sophomores without English AP scores, enrolled after completing English 15 as Freshmen	1967	16		619	30	81.1	6
	1966	16		615	29	81.3	4
French 41, Introduction to French Literature: Seventeenth to Twentieth Centuries							
(a) Freshmen with French AP scores, taking this as first college French course	1967	15	La 2-5 Li 1-4	665	31	80.6	5
(b) Sophomores without French AP scores, taking this after lower-level Freshman French courses	1967	15		660	31	81.1	4
	1966	15		660	34	83.8	4
Mathematics 20, Intermediate Analytic Geometry and Calculus							
(a) Freshmen with mathematics AP scores taking this as first Yale mathematics course	1967	21	3-5	702	33	81.8	9
(b) Sophomores without mathematics AP scores, enrolled after completing mathematics 10 and 15 as Freshmen	1967	21		701	33	76.9	14
	1966	21		696	36	78.3	8

* The broad range of grades in Mathematics (60-95; 50-100; and 60-95 respectively) combined with flatness of the distributions, accounts for the seemingly atypical standard deviations.

non-AP sophomores of comparable ability, in the same courses, indicate that the two groups achieved rather similarly. The AP Program appears to prepare able students to compete adequately with more advanced college students.

While some might wish for greater differences between means, it would be a remote possibility to secure highly significant differences in test data based on samples within such a restricted range of SAT-V scores (i.e., 500-800). While we wish that the samples had been larger, these restrictions were inherent in the data. Fractionization in AP populations is likely to prevent researchers from achieving as conclusive evidence of the validity of the Program as they would like. Finding equally competent non-AP students with whom to compare this select group will remain difficult.

Because the AP Program is such an important innovation in higher education, we would urge colleges to test its validity by conducting controlled experiments under which students are assigned randomly to courses regardless of their AP scores and their participation in an AP Program. Such a plan would permit first the meaningful analysis and comparison of the results of high-and low-scoring AP students and of non-AP groups, and secondly, the calculation of measures of "statistical significance" which are not appropriate with highly selected samples such as those in this study.

REFERENCES

- Blackmer, A. R. (Chm.) *General education in school and college*. A committee report by members of the faculties of Andover, Exeter, Lawrenceville, Harvard, Princeton and Yale. Cambridge: Harvard University Press, 1952.
- Casserly, P. L. College decisions on advanced placement. College Entrance Examination Board Research and Development Reports. RDR 64-65, No. 15. Princeton: Educational Testing Service, Jan. 1966.
- Cornog, W. H. *College admission with advanced standing*. Final report and summary of the June 1955 evaluating conferences of the school and college study. Gambier: The School and College Study of Admission with Advanced Standing. Kenyon College, 1956.
- Wilcox, E. T. Seven years of advanced placement. *College Board Review*, Fall 1962, 48, 29-34.

SAT AND HIGH SCHOOL AVERAGE PREDICTIONS OF FOUR YEAR COLLEGE ACHIEVEMENT

MARVIN SIEGELMAN
City College of New York

SURPRISINGLY little research has been conducted on the most widely used college selection test, the Scholastic Aptitude Test (SAT) of the College Entrance Examination Board (Zimmerman, 1965). Although the SAT has been in existence for over 20 years and is currently used in over 500 colleges and universities, almost all of the validity data have been reported by the Educational Testing Service (Whitla, 1965; Zimmerman, 1965). Most reports also note only the relation between SAT and Freshman Grade Point Average (GPA), ignoring the remaining three years (Fricke, 1965). The contribution that SAT makes to High School Average (HSA) in predicting GPA is another infrequently examined area. The purpose of the present research was to analyze the degree of association between achievement (GPA) during four years of college attendance and (a) SAT Verbal, Mathematics, and Total Scores, (b) HSA, and (c) a composite of SAT and HSA scores.

Method—Subjects. The subjects in the present study consisted of students who had completed four years of study, approximately 128 credits, and graduated from the City College of New York (CCNY). Sat, HSA, and GPA data were collected from the permanent records in the CCNY Registrars office for 80 males and 95 females. Although all students were interested in teaching, they represented all areas of undergraduate concentration in liberal arts and fine arts. The interest in teaching was indicated by their completing a sequence of educational courses that are required for teacher licensing in New York City. Most students came from an upper lower or middle class socio-economic background. The

average age at entering CCNY for males was 17.98 ($SD = 2.06$), and for females it was ($SD = 1.87$).

The admission criteria at CCNY consisted of an Entrance Composite Score (ECS) which included the HSA plus a converted SAT Verbal (SAT-V) and SAT Math (SAT-M) score. In terms of a priori weights, the HSA was intended to contribute 50 per cent to the ECS and the SAT-V plus SAT-M the remaining 50 per cent. The males had the following means (M) and standard deviations (SD) in their scores when they entered CCNY: HSA, $M = 83.86$, $SD = 5.50$; SAT-V, $M = 502.94$, $SD = 83.39$; SAT-M, $M = 529.70$, $SD = 75.69$; SAT: V + M, $M = 1032.64$; $SD = 186.67$; ECS, $M = 167.76$, $SD = 8.81$. When the females entered CCNY their corresponding statistics for their scores were: HSA, $M = 86.46$, $SD = 4.34$; SAT-V, $M = 503.76$, $SD = 89.57$; SAT-M, $M = 504.89$, $SD = 89.74$; SAT: V + M, $M = 1008.65$, $SD = 153.63$; ECS, $M = 171.28$, $SD = 23.42$.

Results and discussion. The most striking aspect of Table 1 is the exceptionally low correlations between SAT and grades for males in contrast to females. The strength of relationship for male SAT findings, but not for the Female results, is also generally lower than that for the typical SAT validity coefficients for male or fe-

TABLE 1
*Correlations between SAT, HSA, and College Grades for Males and Females**

	Males (<i>N</i> = 80)				Females (<i>N</i> = 216)		
Non-cumulative GPA	SAT-V	SAT-M	HSA	ECS	SAT-V	SAT-M	HSA
Upper freshman	-.01	-.05	.25	.21	.27	.29	.31
Upper sophomore	.20	-.09	.25	.20	.38	.31	.30
Upper junior	.01	.09	.39	.35	.29	.33	.36
Upper senior	.05	.16	.25	.29	.24	.33	.33
Cumulative GPA							
Lower freshman	.04	.00	.24	.21	.16	.22	.32
Upper freshman	.00	-.07	.25	.20	.27	.29	.31
Lower sophomore	.06	-.11	.27	.22	.30	.29	.34
Upper sophomore	.10	-.08	.29	.23	.37	.33	.37
Lower junior	.11	-.08	.36	.29	.32	.35	.40
Upper junior	.24	.04	.33	.35	.37	.38	.41
Lower senior	.08	-.02	.43	.37	.39	.39	.41
Upper senior	.10	.03	.44	.40	.39	.38	.41
English average	.16	.06	.23	.29	.29	.15	.22
Science average	.07	-.01	.15	.12	.29	.36	.14
Education average	.06	.08	.20	.20	.28	.28	.29

* Decimal points omitted.

male college freshman GPA, which usually range from .16 to .61 (Zimmerman, 1965).

Greater homogeneity in ability of CCNY students, especially of those who complete four years of college, may account in part for lowered validity coefficients, but the close to zero correlations for males probably can not be attributed entirely to this reduced variability. One could speculate, for example, that the males were less conforming than the females and tended to perform more according to their interests and motivations than according to their academic potential as estimated by the SAT. The correlations between HSA and GPA for males, furthermore, do not indicate especially uniform ability. The efficiency of the SAT for predicting GPA for males, at least at CCNY, must be seriously questioned.

The relationship of HSA to GPA, on the other hand, holds up reasonably well for both males and females. It is comparable to that found in the data reported by Passons (1967) and somewhat lower than that indicated by the Lins, Abell, and Hutchins (1966) findings. For males, the use of the SAT-V plus SAT-M with HSA actually lowered the validity coefficients in the ECS from the predictions made from the HSA alone, whereas for the females there was an increase in ECS over HSA used alone. The median ECS correlation increase for females was .12.

The predictions of cumulative GPA indicate that the accuracy of the HSA and SAT predictors improves for females as the number of completed credits represented in GPA increases. The expanded GPA reliability might be expected to account in part for the higher correlations.

The findings of the present study, especially for males, call attention to the need for additional research on the value of using the SAT as a criteria for admission to college. The expense, concern, and effort involved in securing SAT scores should be reconsidered in terms of how much the SAT contributes to the HSA in predicting GPA.

REFERENCES

- Fricke, B. J. College Entrance Examination Board Admission Testing Program. In O. K. Buros (Ed.), *The sixth mental measurement yearbook*. Highland Park, N.J.: Gryphon Press, 1965. P. 760.

- Lins, L. J., Abell, A. P., and Hutchins, H. C. Relative usefulness in predicting academic success of the ACT, the SAT, and some other variables. *Journal of Experimental Education*, 1966, 35, 1-29.
- Passons, W. R. Predictive validities of the ACT, SAT and High School grades for first semester GPA and freshman courses. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 1143-1144.
- Whitla, D. K. College Entrance Examination Board Admissions Testing Program. In O. K. Buros (Ed.), *The sixth mental measurement yearbook*. Highland Park, N.J.: Gryphon Press, 1965. P. 760.
- Zimmerman, W. S. College Entrance Examination Board Scholastic Aptitude Test. In O. K. Buros (Ed.), *The sixth mental measurement yearbook*. Highland Park, N.J.: Gryphon Press, 1965. P. 449.

THE RELATIONSHIP BETWEEN EXPECTED GRADES AND STUDENTS' EVALUATIONS OF THEIR INSTRUCTORS

DAVID S. HOLMES¹

University of Kansas, Lawrence

THERE has been some concern about the validity of student ratings of their instructors' teaching performances because the students may not be completely objective observers. It has been suggested that the grades the students expect may influence their evaluations of the instructors. Unfortunately, the research on this question (e.g., Anikeef, 1953; Heilman and Armentraut, 1936; Hudelson, 1951; Riley, Ryan, and Lifshitz, 1950; Vocks and French, 1960; Weaver, 1960) has been relatively meager, productive of conflicting conclusions, uninformative because different elements of the evaluation were not considered separately, and in many cases the research was methodologically inadequate. In view of the growing use and importance of student evaluations (Eble, 1970), the present study was carried out to assess the relationship which expected grades have to evaluations.

Method. The evaluation instrument used was the *Teaching Assessment Blank* (TAB) (Holmes, 1971). The TAB's which provided the data for the present study were from seven of the 322 classes surveyed in the College of Arts and Sciences at the University of Texas (Austin). All classes used were lecture classes with enrollments over 100. A more complete description of the classes

¹The present research was carried out during the author's tenure as Research Consultant to the Measurement and Evaluation Center, University of Texas (Austin). The author would like to thank Caroline Dowell, Nancy Earl, Cheryl Lansing, and E. Femme for their assistance. Publication of the report was sponsored by the Educational Testing Service, Princeton, New Jersey.

was presented earlier (Holmes, 1971). In this analysis, however, data from subjects who expected grades of D or F were not considered because combined they constituted fewer than seven per cent of the students. The numbers of students expecting grades of A, B, and C were: 200, 752, and 587, respectively. The analyses were limited to the 18 items which make up the *Instructor Presentation*, *Evaluation-Interaction*, and *Student Stimulation* subscales of the TAB.

Results and discussion. The responses of students expecting A's, B's, and C's were compared on each of the 18 evaluation items in each of the seven classes. The probability values associated with the resulting *F* values are presented in Table 1. If the probability value associated with an item was .10 or less the item was considered to be *related to expected grades* in that class and is italicized in Table 1. This leniency in terms of probability

TABLE 1

Probability Values Associated with F's Arrived at in Comparing Responses of Students' Expected Grades of A, B, and C

Item	A	B	C	Class D	E	F	G
<i>Student Stimulation Subscale</i>							
*11.	.03	.00	.18	.60	.00	.92	.11
**12.	.00	.09	.00	.16	.00	.23	.53
*15.	.02	.00	.18	.71	.04	.87	.57
**24.	.00	.00	.09	.04	.00	.86	.35
**25.	.06	.02	.00	.50	.00	.77	.61
**29.	.00	.01	.00	.00	.01	.01	.29
**30.	.00	.14	.01	.00	.00	.15	.56
<i>Interaction-Evaluation Subscale</i>							
*13.	.00	.62	.02	.32	.01	.23	.72
14.	.01	.80	.11	.54	.01	.96	.20
*19.	.84	.01	.06	.06	.65	.89	.33
**20.	.00	.10	.04	.00	.00	.23	.59
**21.	.00	.18	.04	.00	.00	.51	.63
23.	.03	.07	.53	.33	.25	.31	.90
<i>Instructor's Presentation Subscale</i>							
7.	.76	.23	.81	.09	.11	.87	.80
8.	.02	.11	.34	.11	.00	.81	.04
9.	.10	.03	.26	.51	.00	.53	.63
*10.	.06	.70	.00	.61	.08	.52	.12
16.	.11	.01	.22	.61	.91	.80	.54

* Highly consistent.

** Moderately consistent.

level was compensated for by the demand for replication across classes which will be considered later. In 57 (45 per cent) cases the probability values indicated that the item was related to expected grades. Since fewer than 12 of these 57 values would be expected by chance, it is clear that expected grades were related to evaluation responses. More important, however, were the answers to questions of (a) whether the relationships were consistent across classes, (b) whether the associations were limited to any specific subset of items, and (c) whether the correlations were of practical as well as statistical significance.

Consistency across groups. From inspection of the values presented in Table 1 it is clear that while expected grades were related to the responses given to a number of items in classes A through E, the expected grades were for the most part not related to the responses of the students in classes F and G. In fact, classes A through E averaged 11.2 related items, whereas there was only one related item in each of classes F and G. It is, therefore, apparent that there were clear and important intergroup differences in terms of whether or not expected grades were related to evaluation responses. The question of the "boundary conditions" which determined whether expected grades influenced evaluations will be discussed later. Because it is clear that classes F and G were beyond the limits of prediction, in future analyses only the data from classes A through E will be considered.

In the present study seven items were found to be related to expected grades in four of the five classes being considered. These items will be referred to as *highly consistent* in their relationships to expected grades. Another group of five items was found to be related in three of the five classes, and these items were classed as *moderately consistent* in their relationship to expected grades. These two groups of items accounted for 82 per cent of the italicized probabilities for classes A-E noted in Table 1. Because of the probability level used, most of the remaining italicized probabilities could be attributed to chance. It, therefore, can be concluded that there were a number of items that were consistently related to expected grades.

Content of consistently related items. All items making up the *Student Stimulation* subscale were consistently (five highly, two

moderately) related to expected grades. Students who expected higher grades reported that they paid more attention (No. 11), felt more challenged (No. 12), were more stimulated (No. 15), were more interested (No. 24), looked forward to attending class more (No. 25), made more of an effort to learn (No. 29), and learned more (No. 30) than did students who expected lower grades. The responses to four of the items from the *Interaction-Evaluation* subscale showed a consistent (two high, two moderate) relationship to expected grades. The analyses indicated that students who expected higher grades reported the instructor had more adequate evidence on which to base their grades (No. 20), the grading system was fairer (No. 21), they felt freer to ask questions and disagree (No. 13), and they thought that the instructor returned assignments more promptly (No. 19) than did students who expected lower grades. Interestingly enough, the students' evaluations of the instructors' fairness in dealing with students apart from grading (No. 14) and the instructors' interest in students (No. 23) were not found to be related to expected grades. In sharp contrast to the items measuring the degree to which students were stimulated and their evaluation of the grading process are the items on the *Instructor Presentation* subscale, for none of these was highly consistent in its relationship to expected grades while only an item measuring the degree to which the student felt the instructor was aware of whether the class was following him (No. 10) was moderately consistent in its relationship to expected grades.

In summary, it appears that students who had expected lower grades reported less personal involvement in the course than did students expecting higher grades but, contrary to what might be expected, students anticipating lower grades were not more critical of the instructors' presentations than were students expecting higher grades. In other words, it seems that when a student expected a low grade he did not become critical of the instructor per se but rather he seemed to accept the blame himself and in effect said, "I didn't do better because I didn't become excited in this class and possibly as a result I didn't work hard enough." The only externalization of blame noted in these results involved the criticism of the evaluation system, but because the classes which were used for this analysis were all large ones (100+

students) in which multiple choice examinations played a predominant role in determining grades, the evaluation system was an easy and very possibly a justifiable target for criticism by students expecting low grades.

It cannot necessarily be assumed from these data that the lowered scores on the *Student Stimulation* subscale items of students who had expected low grades resulted from a defensive distortion which they used to explain their relatively poor grades. It may be that students who had anticipated higher grades were more intelligent than those who had expected lower grades, and because of superior ability they were able to benefit more from and be stimulated more by the instructor. With regard to this possibility, it should be noted that in every class there was a significant difference in the actual grade point averages of students expecting A's, B's, and C's. If ability influenced both expected grades and the degree to which a student was stimulated, the relationship between expected grades and evaluation responses would not be a threat to the validity of the responses but would instead indicate that students of different levels of ability were differentially influenced by the instructor. Whether this circumstance is the case generally will have to await further research.

Practical importance of relationships noted. While the above results are of statistical significance, it must be asked whether these relationships might have any practical importance in terms of the potential influence they exert on evaluation items. To answer this concern, within each class, correlations were computed between expected grades and the responses to the seven items which were found to be highly consistent in their relationships to expected grades. Squaring these values indicated the amount of shared variance. The mean amount of variance shared by expected grades and responses to consistently related evaluation items was only five per cent. In only four of the 35 cases was more than 10 per cent of the variance shared, and in no case was more than 13 per cent shared. From these results it is clear that while there was a subset of items which were consistently related to expected grades, the potential influence of this relationship was negligible.

Boundary conditions. A number of variables (e.g., overall evaluation and expected grades) were used to compare the classes in which the evaluation responses were and were not related to ex-

pected grades. Since none of these comparisons yielded interpretable results, it is not possible at the present time to specify the characteristics of classes in which the relationships will and will not be found. However, the above suggestion that the relationship between evaluation responses and expected grades might be mediated by the students' ability offers one possibility. It may be that in those classes in which no relationships were found, the instructors had directed their lecture material at the C level student and, therefore, given this low level of approach, the instructors' presentations might not have differentially stimulated students of different ability levels.

Conclusions. In most classes relationships existed between the grades students expected and the degree to which they reported they were stimulated by the instructor and the degree to which they felt the grading system was fair. On the other hand, items assessing the instructors' presentations were not found to be related to the expected grades. Thus it did not appear that expected grades were related to a general halo effect. Unfortunately, these data did not indicate whether the lower levels of stimulation reported by students anticipating lower grades was a defensive reaction or whether students expecting higher grades were more intelligent and, therefore, were in a better position to be stimulated by the instructors' presentations. While the nature and antecedents of the observed relationships were of considerable theoretical interest in terms of both the evaluation process and the possible differential effect of instructors on different types of students, it is clear that the relationships did not severely distort the evaluation process and therefore did not pose a serious threat to the validity of the evaluation system.

REFERENCES

- Anikeef, A. Factors affecting student evaluation of college faculty members. *Journal of Applied Psychology*, 1953, 37, 458-460.
- Eble, K. *The recognition and evaluation of teaching*. Washington, D.C.: American Association of University Professors, 1970.
- Heilman, J. and Armentraut, W. The rating of college teachers on ten traits by their students. *Journal of Educational Psychology*, 1936, 27, 197-216.
- Holmes, D. The Teaching Assessment Blank (TAB); a form for the student assessment of college instructors. *Journal of Experimental Education*, 1971, 39, 34-38.

- Hudelson, E. The validity of student ratings of instructors. *School and Society*, 1951, 73, 265-266.
- Riley, J., Ryan, B., and Lifshitz M. *The student looks at his teacher*. New Brunswick, N.J.: Rutgers University Press, 1950.
- Voeks, V. and French, G. Are student-ratings of teachers affected by grades? *Journal of Higher Education*, 1960, 31, 330-334.
- Weaver, C. Instructor rating by college students. *Journal of Educational Psychology*, 1960, 51, 21-25.

FACTOR ANALYSIS OF 1970-71 VERSION OF THE COMPARATIVE GUIDANCE AND PLACEMENT BATTERY

JOSEPH GRIMALDI

Marymount College, Tarrytown, New York

EUGENE LOVELESS, JAMES HENNESSY, AND JOHN PRIOR

Queensborough Community College

The City University of New York

THE Comparative Guidance and Placement Program (CGP) is a multi-purpose battery published by the College Entrance Examination Board (CEEb). The battery, which includes biographical, interest, ability and achievement measures, was designed primarily for use at community and two-year colleges. Lunneborg, Greenmun, and Lunneborg, (1970) reported a factor analysis of the 1967 version of CGP. However, the 1970-71 version of the CGP differs in composition from the 1967 version. Consequently, previous factor studies are somewhat outdated.

Purpose and procedure. The 1970-71 version of the CGP battery served as the basis of the present factor analytic investigation. The battery was administered to the freshman class of Queensborough Community College entering in the Fall of 1970 ($N = 1637$). The battery was scored by Educational Testing Service (ETS). The 11 interests scales and the eight cognitive scales were then factor analyzed through utilizing a principal components solution followed by rotation to the Varimax criterion. Unities were placed in the diagonal cells of the test intercorrelation matrix. All principal components that had associated eigenvalues equal to or greater than one were retained for rotation.

Results. Table 1 contains the test intercorrelation matrix. Decimal points have been omitted.

Table 2 furnishes the rotated factor solution. Communalities as well as eigenvalues have been reported. Decimal points have been omitted except for the eigenvalues. The six-factor solution accounted for 70 per cent of the total variance. Factor I was primarily defined by the cognitive scales with the exception of Mosaics. Factors II, III, and IV were primarily (although not completely) defined by interest scales. Biology, Health, and Physical Sciences defined Factor II; Secretarial, Business, Home Economics, and Academic Motivation were loaded on Factor III; and Mathematics, Physical Sciences, Engineering Technology and Social Sciences were weighted on Factor IV. Factor V was also described by cognitive scales—specifically, Mosaics, Letter Groups, Mathematics, and the Year 2000. The remaining interest scales defined Factor VI.

Discussion. Factor I has been interpreted as a scholastic aptitude factor. It apparently encompassed the ability to cope with school tasks (i.e., reading, vocabulary, sentences, mathematics) as well as the ability to follow verbal directions and solve a problem (Year 2000). The Letter Groups test also was loaded on this factor but at a slightly lower level (.45). The latter finding is consistent, since the Letter Groups test is a more nearly pure measure of general reasoning ability than are the other tests that were loaded on this factor. Hence the Letter Groups test was less susceptible to scholastic influences than were other measures.

Factor II was primarily defined by the Biology and Health interest scales. It seemed to represent both theoretical and applied aspects of the Biology-Health domain. Factor III was primarily defined by the Secretarial and Business interest scales and by the Academic Motivation scale. This pattern of loadings suggested an interpretation of this factor, viz., Practical Interest. The factor bore a similarity to the Practical Outlook scale of the Omnibus Personality Inventory (Heist and Yonge, 1968). Factor IV was primarily defined by the Engineering Technology interest scale. An interpretation of this factor as Interest in Technological Science would appear consistent with the presence of the other variables that also were weighted on this factor, viz., Physical Science and Mathematics interest scales.

TABLE 2

Primary Factor Loadings of CGP Variables for 1970 QCC Freshman Class Using the Varimax Solution

Variable	I	II	III	Factor IV	V	VI	R ²
Interest Scales							
Math	-.01	.06	.31	.69	-.19	-.11	.63
Phy. Sc.	.05	.42	-.08	.79	.13	.12	.83
Eng.	-.06	.00	-.06	.78	-.06	.25	.69
Bio.	.09	.84	-.03	.26	.06	.23	.83
Health	-.02	.89	.08	.06	.01	.07	.80
Home Ec.	.01	.43	.56	-.27	-.17	.36	.73
Sect.	-.06	-.02	.90	-.04	-.01	.05	.81
Bus.	.01	-.06	.84	.28	.19	.05	.83
Soc. Sc.	.31	-.02	.25	.31	.38	.37	.53
Fine Arts	.03	.19	.13	.05	-.16	.83	.76
Music	.05	.09	-.01	.13	.09	.78	.65
Cognitive Scales							
Read.	.90	.02	-.07	.03	.02	.08	.82
Verbal	.91	.03	-.08	.01	.02	.09	.85
Sentences	.74	.07	.05	-.21	-.27	.08	.68
Math	.63	-.03	-.06	.29	-.43	-.15	.70
Yr. 2000	.68	-.05	-.05	.02	-.40	-.02	.62
Mosaics	.15	-.03	-.01	.11	-.73	.07	.57
Lett. Grps.	.45	.01	.04	-.03	-.65	.00	.63
Academic Motivation	-.28	.27	.39	.00	-.11	-.01	.32
Eigenvalues	3.96	3.28	1.99	1.63	1.36	1.05	13.27
Per Cent of Variance	21	17	11	08	07	06	70

Factor V appeared to be a combination of measures. It was primarily defined by tasks of nonverbal reasoning nature (e.g. Mathematics, Year 2000, and Letter Groups) as well as by a measure of perceptual efficiency (Mosaics). Tentatively, this factor might be defined as a Perceptual-Reasoning factor. However, the interrelationship of the tests that were loaded on Factor V as well as their relationship to external criteria bears further investigation.

In summary, the CGP battery was found to yield six interpretable factors. Two of these factors were related to the cognitive tests and four were related to the interest measures. The pattern of loadings split very neatly according to content (i.e., interest vs. cognitive). It would, therefore, seem more profitable if future factor analytic studies divided the battery into portions prior to factoring. The only measure in the battery that could not unequivocally be placed in one or the other category was the Aca-

ademic Motivation scale. Therefore, future studies should include the Academic Motivation scale with both groups of tests.

REFERENCES

- Heist, P. and Yonge, G. *Omnibus Personality Inventory: Form F*, Manual. New York: The Psychological Corporation, 1968.
- Lunneborg, C. E., Greenmum, R., and Lunneborg, P. W. A factor analysis of the core elements of the CEEB Comparative Guidance and Placement Battery. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1970, 30, 387-392.

CORRELATES OF A PASS-FAIL DECISION FOR ADMISSION TO CANDIDACY IN A DOCTORAL PROGRAM IN EDUCATION

WILLIAM B. MICHAEL, ROBERT A. JONES, HUDHAIL AL-AMIR,
CALVIN M. PULLIAS, MICHEL JACKSON, AND VALERIE GOO

University of Southern California

PRIOR to admission to candidacy in the doctoral program of the School of Education at the University of Southern California (USC) students have been required not only to complete the aptitude portion of the Graduate Record Examination (GRE) but also to take the Comprehensive Examination (CE), a battery of five objective examinations—one two-hour test in each of the three fields of Administration, Psychology, and Social and Philosophical Foundations, and two one-hour tests in Curriculum (Elementary Education and Secondary Education or Elementary Education and Higher Education or Secondary Education and Higher Education). Within two to four weeks after a student completes the CE, the Doctoral Committee consisting of approximately ten full-time faculty members reviews cumulative folders; evaluates examination data, prior grades, letters of recommendation, and professional experience and objectives; interviews the candidate; and renders a pass or fail decision regarding his readiness to be admitted to the doctoral program.

Purpose. The purpose of the investigation was to explore the interrelationships among a number of selected variables involved in the Doctoral Committee's pass-fail decision for a sample of 844 students (694 men and 150 women) at USC who took the CE from July 1968 to October 1970 (a period which covered thirteen administrations of the CE), and of whom 793 appeared before the Doctoral Committee. Such information was judged

TABLE 1

Intercorrelations of Comprehensive Examination Scores, Graduate Record Examination Scores, Sex Membership, and the Dichotomous Pass-Fail Decision of the Doctoral Committee Regarding Admission to Candidacy in the Doctoral Program in the School of Education at the University of Southern California for the Period July 1968 to October 1970^a

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1. Comprehensive Examination—Administration	—	.844	.844	.299	.766	.623	.844	.844	.613	.613	.613	.799	.793
2. Comprehensive Examination—Psychology	.39	—	.844	.299	.766	.623	.844	.844	.613	.613	.613	.799	.793
3. Comprehensive Examination—Social and Philosophical Foundations	.33	.57	—	.299	.766	.623	.844	.844	.613	.613	.613	.799	.793
4. Comprehensive Examination—Elementary Education	.48	.39	.32	—	.221	.78	.299	.299	.207	.207	.207	.287	.285
5. Comprehensive Examination—Secondary Education	.51	.46	.40	.46	—	.545	.766	.766	.553	.553	.553	.734	.719
6. Comprehensive Examination—Higher Education	.52	.42	.43	.45	.48	—	.623	.623	.466	.466	.466	.557	.582
7. Comprehensive Examination—Curriculum: (4) + (5) or (4) + (6) or (5) + (6)	.59	.50	.47	.86	.85	.85	—	.844	.613	.613	.613	.799	.793
8. Comprehensive Examination—Total Score	.74	.76	.72	.72	.76	.88	.21	.27	—	.613	.613	.799	.793
9. Graduate Record Examination—Quantitative Score	.18	.25	.23	.18	.20	.17	.21	.27	.35	—	.613	.582	.573
10. Graduate Record Examination—Verbal Score	.11	.49	.48	.22	.27	.24	.29	.42	.81	.82	—	.582	.573
11. Graduate Record Examination—Total Score	.18	.45	.43	.24	.28	.25	.30	.42	.23	.13	.07	—	.754
12. Sex ^b	—	.12	.08	—	.04	—	.02	.00	.23	.29	.27	.01	—
13. Doctoral Committee Decision—Pass or Fail ^c	.53	.55	.50	.49	.51	.53	.59	.70	.16	.29	.27	.01	—

^a Correlation coefficients are below the diagonal, and corresponding sizes of samples are above the diagonal. For example, the correlation between variables 6 and 12 of -.02 found in the twelfth row and the sixth column is based on a sample size of 587, which is cited in the sixth row and twelfth column.

^b For variable 12, the value of 2 was assigned to males and 1 to females.

^c For variable 13, the value of 2 was assigned to "pass" and 1 to "fail."

to be of considerable importance in evaluating the validity of examinations in the admission of students to candidacy.

Findings. Calculated through use of a computer program to allow for missing data for a number of students, the zero order coefficients of correlation in Table 1 point to the following major findings: (1) Moderate intercorrelations among the five parts of the CE varied from .32 to .57. (2) Low to moderate validity coefficients ranging from .11 to .49 were registered for the part and total scores of the GRE. With respect to both separate and total CE scores, higher validities appeared for the Verbal (V) than for the Quantitative (Q) scores, the one exception being for the CE in Administration. Higher validities were found for the GRE scores relative to the foundation fields of Psychology and of Sociology and Philosophy than relative to applied fields of Curriculum and Administration. (3) With respect to the pass-fail criterion part scores of the CE showed validities ranging from .49 to .59; and the total score, a coefficient of .70. (4) The GRE total and part scores exhibited negligible to slight correlations with the pass-fail criterion as evidenced by coefficients varying between -.13 and .27. (5) Sex membership revealed negligible correlation coefficients with CE scores as well as with GRE scores and a coefficient of only .01 with the pass-fail criterion.

Conclusions. For a large sample of 844 graduate students seeking admission over slightly more than a two-year period as candidates for the doctorate in the USC School of Education, part and total scores on the GRE were negligibly to modestly related to performance on five objective achievement tests comprising the Comprehensive Examination (CE). Admission to candidacy based on a pass-fail decision of a doctoral committee was substantially dependent on total scores on the CE, slightly or negligibly related to GRE scores, and virtually uncorrelated with the sex of the candidate.

APPROPRIATENESS OF SUBTESTS IN ACHIEVEMENT TESTS SELECTION

THOMAS M. GOOLSBY, JR.

University of Georgia

AUTHORS and publishers of extensive achievement test batteries have usually presented validity evidence of "content" and/or "curricular" types. This validity evidence has been quite appropriate, since a criterion considered to be better than the battery and/or its subtests would be extremely difficult to identify or define.

Considerations of relationships between total scores and interrelationships among subtests for any two achievement batteries should be given particular attention by those responsible for achievement test selection. For extensive achievement batteries, correlations of total test scores on any two will approach .90 and many times exceed that value. It is possible for the zero order relationship of total scores on two batteries to approach unity (1.00) when one considers the very high reliabilities of batteries approximating .95. On the basis of the .90 magnitude of relationship for total scores, one might decide that one battery is just as appropriate to use with a given population as another. This decision is probably correct if one is not particularly interested in determining subject area emphases for follow-up activities or curriculum guidance values of different achievement test batteries. Users of achievement test batteries are generally more interested in follow-up activities and in curriculum guidance features of achievement tests than in a simple rank ordering or placement of students on the basis of a total score. The present study presents evidence of internal uniqueness of two achievement test batteries important to diagnosis and guidance.

The present study was designed to

1. determine the degree to which like named components (science, mathematics and so forth) of two achievement test batteries measure the same things.
2. ascertain the correlation of total test scores for the two achievement test batteries.
3. derive some information regarding the rational appropriateness of achievement test batteries for a given population

Procedures. The two achievement test batteries considered were the Metropolitan Achievement Tests (MAT) and Stanford Achievement Tests (SAT) designed for measurement of outcomes of instruction at the same grade level. Selected subtests which are named exactly the same in both batteries and others which are similarly named were considered.

The population used was a large junior high school in a suburban area of greater metropolitan Miami. The MAT and SAT were administered to the seventh grade students in the fall. Intercorrelations among and between certain of the measures and between total scores for the two batteries were obtained. Means and standard deviations were computed for certain subtests of the two batteries.

Results and discussion. Total scores on the MAT and SAT correlated .89. At this level of relationship, it can be said that the two batteries measure essentially the same characteristics. This outcome was expected. Even though these two batteries apparently measure similar components, the evidence that follows shows marked internal and practical differences important to achievement testing.

Table 1 presents relationships between certain subtests of MAT and SAT. The relationships corrected for attenuation between similarly named subtests in MAT with those in SAT had a range of .49 to .69. The differently named subtests had a range of .39 to .57. These ranges are, in general, typical of interrelationships corrected for attenuation for subtests in a single achievement battery such as MAT or SAT. Those subtests in a single battery for which this range is not typical are subtests in a battery for special purposes such as vocabulary or special subject area coverage known to be highly related to other components but deemed important to include for special curricular emphases.

The relationships in Table 1 for similarly named subtests in the

TABLE 1

*Relationships between Certain Subtests of MAT and SAT (N = 335)**

Variables ^b	1	2	3	4	5
6	53 ^c (58) ^d	46 (51)	46 (54)	40 (44)	48 (52)
7	45 (49)	52 (58)	52 (62)	43 (48)	50 (55)
8	39 (44)	48 (55)	52 (64)	53 (61)	58 (66)
9	39 (44)	44 (50)	47 (51)	50 (56)	59 (66)
10	49 (56)	55 (64)	57 (70)	49 (57)	60 (69)

* Decimals omitted.

^b 1 SAT-Arithmetic Computation.

2 SAT-Arithmetic Concepts.

3 SAT-Arithmetic Application.

4 SAT-Science.

5 SAT-Social Studies.

6 MAT-Arithmetic Computation.

7 MAT-Arithmetic Problem Solving.

8 MAT-Science.

9 MAT-Social Studies Information.

10 MAT-Social Studies Skills.

* Zero order correlations.

^d Zero order correlations corrected for attenuation are inserted within parenthesis.

two batteries are probably somewhat surprising to many of those who frequently deal with practical measurement problems such as achievement battery selection to measure outcomes of instruction, pupil progress, and uniqueness of a given curriculum. The evidence presented here is supportive of the fact of uniqueness of curriculum

TABLE 2

Means and Standard Deviations for Certain MAT and SAT Subtests (N = 335)

Tests	Descriptive Statistics		Means	SD
	k ^a	r _{xx} ^b		
SAT-Arithmetic Computation	41	.87	13.61	5.45
SAT-Arithmetic Concepts	40	.82	15.69	5.36
SAT-Arithmetic Application	36	.77	11.17	4.43
SAT-Social Studies	92	.89	40.47	14.01
SAT-Science	60	.88	25.06	8.88
MAT-Arithmetic Computation	45	.91	20.90	8.04
MAT-Arithmetic Problem Solving	48	.92	22.24	8.11
MAT-Science	55	.83	27.08	9.85
MAT-Social Studies Information	60	.89	27.77	9.91
MAT-Social Studies Study Skills	40	.84	19.75	6.76

^a Number of items.^b Corrected split test reliability coefficient estimates reported by the authors of the tests.

TABLE 3

Multiple Correlations Involving Certain MAT and SAT Subtests
($N = 335$)

Dependent Variable ^a	Predictor ^a	R ^b
7	2,3	57
6	2,3	51
2	6,7	52
3	6,7	52
5	9,10	65

^a Variables described as follows: 2—SAT-Arithmetic Concepts; 3—SAT-Arithmetic Application; 5—SAT-Social Studies; 6—MAT-Arithmetic Computation; 7—MAT-Arithmetic Problem Solving; 9—MAT-Social Studies Information; 10—MAT-Social Studies Skills.

^b Decimals omitted.

areas as measured by two achievement batteries designed to measure outcomes at the same grade level. The evidence could be interpreted as being supportive of the need for a variety of batteries constructed to fit different curriculums by subtests in both content and/or method.

Table 2 presents the means and standard deviations for certain for certain of the MAT and SAT subtests. In general, the MAT is more appropriate for the population of this study. The SAT subtests presented are, in general, somewhat too difficult. The reader should not, however, depend too much on these values. The MAT was selected at an earlier date for the population of this study by analysing the curriculum by areas, considering future goals and innovations, and administering subtests of various batteries to appropriate instructional personnel responsible for instruction in a given subject area.

Table 3 presents multiple correlations of certain MAT and SAT subtests. These multiple correlations are not substantially different from the zero order relationships presented in Table 1.

REFERENCES

- Metropolitan Achievement Test. New York: Harcourt, Brace and World, 1959.
Stanford Achievement Test. New York: Harcourt, Brace and World, 1963.

THE RELIABILITY AND VALIDITY OF THE CONTEMPORARY MATHEMATICS TEST

JOSEPH S. RENZULLI

AND

ROBERT A. SHAW

University of Connecticut

THE Contemporary Mathematics Test (CMT) is a relatively new instrument that is designed to measure the extent to which students have mastered course content in modern mathematics. The test series, which consists of two forms at five levels (Lower Elementary, Upper Elementary, Junior High, Senior High, and Algebra), yields total raw scores which can be converted to percentile ranks, standard scores, or stanines through the use of normative tables. According to the Manual, the test series measures results common to programs in contemporary mathematics. Emphasis is placed on understanding concepts and skills pertinent to solving problems in the areas of structure and number as well as on special mathematical devices such as number lines and the coordinate system, formulas, and special symbols (California Test Bureau, 1966a). The content categories for all levels except algebra are defined as properties of numbers, mathematical structures, systems of enumeration, nature and structure of proof, ratio and proportion, mathematical sentences, geometry, and other topics including variables, functions, and graphs (California Test Bureau, 1966b).

Information pertaining to the reliability and validity of the CMT is limited to data gathered on the norming sample by the test publishers. Reliability data of the usual type are reported by form and grade level. Internal-consistency reliability coefficients computed by use of the Kuder-Richardson formulas 20 and

21 range from .70 to .89; short- and long-range stability data computed by use of the Pearson product-moment correlation coefficients (corrected for range) range from .60 to .88. Congruent validity data comparing the CMT (Junior High Level) with the California Short-Form Test of Mental Maturity (Junior High Level, 1957 edition) yielded product-moment correlations ranging between .50 and .62. Comparisons with the California Achievement Test (CAT) (Junior High Level, 1957 edition) yielded correlations ranging from .43 between CMT and CAT Spelling to .70 between CMT and CAT Arithmetic Reasoning.

Purpose. The purpose of this study was to obtain additional empirical information relating to the reliability and validity of the Contemporary Mathematics Test—Junior High Level. Although two reviewers (Smith, 1967; Romberg, 1968) have expressed concern about the lack of empirical support for this instrument, a review of the literature indicates that apparently no independent research studies relating to reliability and validity have been carried out to date.

Procedure. A sample of 232 students in grades seven, eight, and nine who were enrolled in a modern mathematics program for a minimum of three years served as subjects for the present study. Because of incomplete comparative data, relationships between the CMT and other selected variables are based on smaller samples. As a group, the subjects were somewhat above average in general mental ability (Mean IQ = 112); however, the entire spectrum of ability levels was represented in the sample population with the exception of students enrolled in special classes for the mentally retarded.

Reliability data consisted of pre- and post-test comparisons between the two forms of the CMT administered at the beginning and at the end of the school year and of estimates derived from the Kuder-Richardson 20 formula. Validity data consisted of comparisons between each of the two forms of the CMT and scores on (1) a comprehensive final examination that was based on the year's work in mathematics, (2) final grades in mathematics, (3) the Arithmetic Reasoning and Arithmetic Fundamentals subtests of the California Achievement Test, (4) the mathematics portion of the Sequential Test of Educational Progress, and (5) the Lorge-Thorndike Intelligence Test.

Results and discussion. The summary statistics for the above comparisons appear in Tables 1 and 2. As can be seen in Table 1, correlations based on retesting with alternate forms of the CMT ranged from a low of .75 to a high of .82. Correlations derived from the Kuder-Richardson formula 20 were slightly higher, ranging between .78 and .88. Thus, the reliability of two forms of the CMT over a period of approximately one school year appeared to be quite favorable, and reliability estimates based on individual item statistics indicated that the two forms of the CMT possess a relatively high degree of internal consistency.

Data pertaining to the validity of the CMT are presented in Table 2. Both forms of the test seem to be more closely related to STEP scores, IQ scores, course grades, and final examination grades than to subscores on the Arithmetic Subtests of the CAT. It should be noted that the CAT Arithmetic portion showed a lower relationship with these variables than did the CMT. These findings may reflect the current emphasis on teaching mathematical concepts and the focus of the CAT upon traditional skills. The relatively low correlations between the CMT on the one hand and final examinations and course grades on the other suggested that the test is largely independent of the content to which the subjects in the study were exposed. This outcome may be a result of the scope of the test, the emphasis given by teachers to the understanding of modern mathematical concepts, or an interaction between these two factors.

TABLE 1

Means, Standard Deviations and Correlation Coefficients for Pre- and Post-Administrations of the CMT

	N	Administration	Mean	Standard Deviation	KR-20 Correlation Coefficient	Pearson Product Moment Correlation Coefficient
Grade 7	103	Pre*	15.12	7.10	.81	.80
		Post**	17.57	7.94	.84	
Grade 8	96	Pre	17.40	6.91	.80	.81
		Post	21.01	9.17	.88	
Grade 9	33	Pre	24.27	5.76	.78	.75
		Post	26.58	6.24	.84	
Total	232	Pre	17.36	7.45	.81	.82
		Post	20.28	8.77	.87	

Note.—All coefficients are significant at the .001 level.

* Form W.

** Form X.

TABLE 2

Inter-Correlation among CMT (Forms W and X) and Seven Selected Variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1) CMT-W	—	.83**	.59**	.41	.17	.29	.39**	.37**	.77**
N		70	70	33	33	33	70	70	40
2) CMT-X		—	.73**	.17	.22	.23	.37**	.40**	.81*
N			70	33	33	33	70	70	40
3) Lorge-Thorndike			—	.42*	.58**	.48*	.35**	.38**	.75**
N				33	33	33	70	69	38
4) CAT-AR ^a				—	.54**	.76**	.28	.25	.65**
N					33	33	33	33	33
5) CAT-AF ^b					—	.95**	.33	.47*	.48*
N						23	33	33	33
6) CAT-Total						—	.38	.47*	.60**
N							33	33	33
7) Final Examination							—	.75**	.56**
N								69	36
8) Course Grades								—	.45**
N									36
9) STEP									—

* Indicates significance at .05 level.

** Indicates significance at .01 level.

^a California Achievement Test-Arithmetic Reasoning.^b California Achievement Test-Arithmetic Fundamentals.

REFERENCES

- California Test Bureau. *Contemporary Mathematics Test: Manual for administering all tests in CMT series*. Monterey, Calif.: McGraw-Hill, 1966. (a)
- California Test Bureau. *Contemporary Mathematics Test: Technical report*. Monterey, Calif.: McGraw-Hill, 1966. (b)
- Romberg, T. A. Test Review: Contemporary Mathematics Test Series. *Journal of Educational Measurement*, 1968, 5, 349-351.
- Smith, F. M. Test Review: Contemporary Mathematics Test. *Journal of Educational Measurement*, 1967, 4, 123-124.

DIFFERENTIAL EFFECTS OF INITIAL COURSE
PLACEMENT AS A FUNCTION OF ACT MATHEMATICS
SCORES AND HIGH-SCHOOL RANK-IN-CLASS IN
PREDICTING GENERAL PERFORMANCE
IN CHEMISTRY

JOHN R. REINER

Southern Illinois University at Edwardsville

JUSTIFICATION for using standardized tests as a device for placing new college students at suitable beginning levels of instruction generally has centered on the possible effects of poor placement on motivation and performance, or loss of time in attaining educational goals (Dunn, 1966).

Although such placement has become commonplace, validity studies ordinarily examine the individual course in question, and the criterion consequently becomes a measure which reflects the achievement of only one term. An example of this method was a recent validity study of the American College Testing Program's Mathematics Placement Examination (Shevel and Whitney, 1969).

The present study examined the possible long-term effects of initial placement. The specific question addressed was, "What effect does initial placement of students in the beginning course of a discipline have on subsequent performance in the discipline?"

Methodology. Subjects were 250 students who, as entering freshmen, had been placed in the first of a three-course beginning chemistry sequence (normally considered the appropriate beginning course for science majors) and 88 students who had been placed in the second course of the same sequence on the basis of a locally-developed examination. In the course of a validity study of this procedure, it was discovered that the locally-derived instrument was contributing very little to the proper placement of students

$R^2 = 0.13$). Consequently, the same subjects used in the validity study were employed to develop a "best" prediction model for future placement, which resulted in a two-variable equation using the American College Testing Program Standard Mathematics Score (ACTM) and high school rank-in-class percentile (RIC).

At the end of the Spring Quarter 1970, after the originally-placed students had been in attendance for two academic years, their grades in six "key" chemistry courses were recorded, if available, and used in the present study.

Statistical procedures followed a multiple-regression technique to obtain F ratios by which to judge the significance of the extent to which predictor variables account for the variability of a criterion (Kelly, Beggs, and McNeil, 1969). In this study, the criterion was cumulative grade-point-average (GPA) in the six key chemistry courses, and predictors were continuous scores ACTM and RIC and two categorical predictors (PL_1 and PL_2) representing "first course placement" or "second course placement." The formal hypothesis tested was: "The model representing effects of PL across all levels of RIC, across all levels of ACTM, does not add significantly to the predictability of GPA by a model representing the linear effects of the three variables alone." The significance of the effects of restricting the "interaction" model efficiency (R^2) was tested by examining the obtained F in terms of the probability level (α) established for this study, 0.05.

Analysis of results. Data used to test the hypothesis are shown in Table 1. On the basis of these data, the formal hypothesis was rejected and an alternate hypothesis was accepted as tenable: The model representing effects of PL across all levels of RIC, across all levels of ACTM, adds significantly to the predictability of GPA by a

TABLE 1

Test Data for Determining Effectiveness of Differential Predictor Interaction in GPA Prediction

df_f	df_r	R_f^2	R_r^2	F	P
2	332	0.33	0.28	12.98	<0.001

Note.— df_f and df_r are degrees of freedom for the "full" and "restricted" models, respectively. The subscripts "f" and "r" with R^2 have similar meaning. $df_f = (m_1 - m_2)$ where m_1 is the number of linearly independent vectors in the full model and m_2 is the number of linearly independent vectors in the restricted model. $df_r = (N - m_1)$ where N is the number of subjects and m_1 is defined as above.

model representing the linear effects of the three variables alone. To state these results in another way, initial level of placement has differential effects on GPA at different levels of RIC across different levels of ACTM.

The general prediction equation developed to reflect the model tested was as follows:

$$Y_i = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_8X_8, \text{ where:}$$

Y_i = Criterion = GPA

X_1 = Exclusive group membership (PL₁)

X_2 = Exclusive group membership (PL₂)

X_3 = (RIC) (PL₁)

X_4 = (RIC) (PL₂)

X_5 = (ACTM) (PL₁)

X_6 = (ACTM) (PL₂)

X_7 = (RIC) (ACTM) (PL₁)

X_8 = (RIC) (ACTM) (PL₂)

a = constant or regression weight associated with the unit vector.
 b_1 through b_8 = weights associated with X_1 through X_8 and calculated to maximize the variance accounted for by the model.

Using the calculated values the general equation becomes,

$$Y_i = 0.7318 + 2.2929X_1 + 1.0300X_2 - 0.0194X_3 - 0.0290X_4 - 0.0469X_5 - 0.0146X_6 + 0.0014X_7 + 0.0020X_8$$

Then two equations can be developed using the predictors investigated:

Placement in first course:

$$Y_i = 0.7318 + 2.2929(1) + 1.0300(0) - 0.0194(\text{RIC}) - 0.0469(\text{ACTM}) + 0.0014(\text{ACTM})(\text{RIC})$$

Placement in second course:

$$Y_i = 0.7318 + 2.2929(0) + 1.0300(1) - 0.0290(\text{RIC}) - 0.0416(\text{ACTM}) + 0.0020(\text{ACTM})(\text{RIC})$$

Table 2 shows the predicted grade-point-average for students according to the course into which they were initially placed, at representative levels of ACTM-RIC combinations. The data in Table 2 are not absolutely complete, for they are only meant to illustrate a continuous, three-variable interaction. It is clear, however, that initial placement has a differential effect, at different levels of ACTM

TABLE 2

Predicted Grade-Point-Averages at Representative Levels of ACTM-RIC Combination
 (A = 5, B = 4, C = 3, D = 2, E = 1)

ACTM	Placement Level	RIC			
		40	60	80	99
24	1st Course Placement	2.47	2.75	3.04	3.30
	2nd Course Placement	1.52	1.90	2.28	2.64
26	1st Course Placement	2.49	2.83	3.17	3.49
	2nd Course Placement	1.60	2.06	2.53	2.96
28	1st Course Placement	2.50	2.90	3.30	3.67
	2nd Course Placement	1.68	2.22	2.76	3.27
30	1st Course Placement	2.52	2.97	3.43	3.86
	2nd Course Placement	1.75	2.37	3.00	3.58
32	1st Course Placement	2.54	3.05	3.56	4.04
	2nd Course Placement	1.83	2.53	3.24	3.90
34	1st Course Placement	2.56	3.12	3.69	4.22
	2nd Course Placement	1.91	2.69	3.47	4.21

and RIC, on subsequent performance. Comparing the predicted performance of the two groups in the courses in question, it is apparent that students initially placed in the first course would be expected to attain higher GPA's if their RIC's were in the middle or lower ranges, regardless of ACTM. As ACTM becomes greater, of course, the predicted scores for the two groups become more and more nearly similar. It is only in the *highest* ranges of RIC distribution that students initially placed in the second course would be expected to do as well or surpass those initially placed in the first course.

Implications. This study was concerned only with a single discipline at a single institution, and further information about the phenomenon investigated should be gathered by replication at different institutions, and certainly by applying the basic procedures used to data from different disciplines. However, the findings do support at least this important point: the uncritical acceptance of placement by examination as a device to benefit students may not be an unmitigated good. Although the grade-point-average might not provide a very satisfactory *educational* criterion, it is widely used as a *selective* criterion by graduate schools and prospective employers. Thus, any procedure which by virtue of only theoretically valuable edu-

ational devices places students in a less competitive graduate school or employment position, deserves serious and continuing study.

From a practical viewpoint, the use of an equation of the complexity developed in this study is quite possible with a reasonably sophisticated computer installation. Regardless of specific methods, it is important to note that in order to benefit students, placement in advanced courses should be based only on relatively high scores on relevant predictor variables to avoid the problems discussed. It is possible that the positive aspects of motivation and savings of time associated with advanced placement are outweighed by decrements in subsequent performance. In any event, extensive cross-validation efforts would be required before the results of this investigation could be generalized to this population or to other populations.

REFERENCES

- Dunn, J. E. A study of the University of Arkansas Mathematics Entrance Exam as a placement device. *Journal of Experimental Education*, 1966, 34, 62-68.
- Kelly, F. J., Beggs, D. L., and McNeil, K. A. *Research design in the behavioral sciences: Multiple regression approach*. Carbondale and Edwardsville, Illinois: Southern Illinois University Press, 1969.
- Shevel, L. R. and Whitney, D. R. Predictive validity of the mathematics Placement Examination. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 895-901.

THE CRITERION-RELATED VALIDITIES OF COGNITIVE AND NONCOGNITIVE PREDICTORS IN A TRAINING PROGRAM FOR NURSING CANDIDATES

WILLIAM B. MICHAEL, RUSSELL HANEY, AND YOUNG B. LEE
University of Southern California

AND

JOAN J. MICHAEL
California State College, Long Beach

It was the three-fold purpose of this investigation (1) to cite the validity coefficients of seven standardized cognitive test measures, four indices of high school achievement, and two scales from each of two self-report inventories in the prediction of grades in each of eight courses in a program of nursing education for the 1969-1970 period taken by a total sample of 128 students at the Los Angeles County Hospital, (2) to report validity coefficients with respect to the same combinations of predictor and criterion variables just mentioned for a sample of 96 candidates who survived the first part of the program and continued during the second segment, and (3) to indicate for this sample of 96 successful candidates the validity coefficients of the same predictor variable with respect to each of eight additional criterion measures representing other course work in the nursing program. For the first two purposes the findings are cited in Table 1, and for the third purpose the corresponding data are furnished in Table 2. In addition, the intercorrelations within each of the two sets of criterion measures are presented in these tables. Additional information regarding many of the measures employed as well as findings with prior samples may be found in the article by Michael, Haney, and Jones (1966) and in the previous articles cited in its bibliography.

Findings and interpretation. The findings may be summarized

TABLE 1

*The 1969-1970 Nursing Class at the Los Angeles County Hospital: Validity Coefficients of 15 Predictor Variables with Each of Eight Criterion Measures Both for the Total Sample and for the Group of Successful Candidates Including Intercorrelations of the Eight Criterion Measures**

Predictors (Variables 1-15) and Criterion Measures (Variables 16-23)	Total Sample (N = 128) ^b								Sample of Successful Candidates (N = 96) ^c							
	Eight Criterion Variables (16) to (23)								Eight Criterion Variables (16) to (23)							
	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)
1. Cal Rdg Test (Voc)	20	21	21	16	02	29	12	15	23	31	28	12	30	04	18	18
2. Cal Rdg Test (Comp)	34	34	36	31	21	36	26	26	36	35	42	36	38	25	29	29
3. Cal Rdg Test (Total)	31	32	33	28	15	37	23	27	35	47	42	30	40	15	32	32
4. Cal Math Test (Reus.)	18	24	19	18	16	25	27	20	14	19	23	18	15	26	17	17
5. Cal Math Test (Fund)	18	13	11	11	17	13	15	18	11	12	14	07	08	10	08	08
6. Cal Math Test (Total)	19	17	13	14	14	20	19	17	13	17	20	13	14	11	18	13
7. EAST No. 5-Spa. V.	08	10	04	05	11	07	20	08	06	07	08	08	20	03	17	05
8. High School (HS) GPA	24	23	14	21	23	18	12	16	17	14	12	14	11	11	02	14
9. HS GPA in Solids	28	23	15	26	27	23	10	20	22	16	13	20	18	18	03	16
10. HS Science GPA	20	14	02	14	22	15	05	15	14	06	-01	06	12	10	00	13
11. HS Chem GPA	18	11	10	18	23	21	16	13	09	03	04	12	17	19	03	07
12. 16 PF-Factor G	-04	-03	-05	-02	00	00	-11	-13	-12	-18	-12	-15	-13	-12	-12	-23
13. 16 PF-Factor Q ₂	13	06	21	02	08	15	08	24	09	05	25	05	10	08	05	29
14. MMPI-Mf	07	08	13	22	12	25	19	22	00	00	02	20	-11	18	-20	-14
15. MMPI-Sc	-14	-16	-09	-17	-10	-09	-03	-13	-18	-25	-18	-23	-24	-20	-04	-18
16. Anatomy	—	60	50	49	52	53	41	48	—	55	47	39	43	31	31	36
17. Introd. to Prof. Nursing	60	—	41	51	46	58	39	46	55	—	39	45	45	24	24	39
18. Nursing 1A	50	48	—	62	45	62	43	45	47	39	—	54	46	27	21	33
19. Nursing 1B	49	51	62	—	56	68	43	60	39	45	54	—	54	52	26	33
20. Nutrition	52	46	45	56	—	61	41	62	43	45	46	54	—	51	36	58
21. Pharmacology 1A	53	58	62	68	61	—	45	61	31	24	27	52	51	—	17	48
22. Physiology	41	39	43	43	41	45	—	45	31	24	21	26	36	17	—	35
23. Psychology	48	46	45	60	62	61	45	—	36	39	33	33	58	48	35	—

* All decimal points omitted from the correlation coefficients.

^b Coefficients of .17 and .23 are significant at the .05 and .01 levels, respectively.

^c Coefficients of .20 and .26 are significant at the .05 and .01 levels, respectively.

TABLE 2

Validity Coefficients of 15 Predictor Variables with Each of Eight Additional Criterion Measures for the Group of Successful Nursing Candidates Including Intercorrelations of the Eight Additional Criterion Measures ($N = 98$)^{a,b}

Predictors (Variables 1-15) and Criterion Measures (Variables 24-31)	Eight Additional Criterion Variables (24) to (31)							
	(24)	(25)	(26)	(27)	(28)	(29)	(30)	(31)
1. Cal Rdg Test (Voc)	31	-.09	14	29	12	-10	-.02	-.10
2. Cal Rdg Test (Comp)	24	.07	19	46	23	00	-.13	-.09
3. Cal Rdg Test (Total)	29	.01	20	45	22	-.04	-.10	-.10
4. Cal Math Test (Reas.)	11	.00	09	24	23	30	-.06	.09
5. Cal Math Test (Fund)	01	.00	09	14	00	14	-.03	.16
6. Cal Math Test (Total)	06	.00	10	20	08	23	-.05	.14
7. EAST No. 5-Spa. V ₂	08	.04	-.09	23	-.11	13	.00	.00
8. High School (HS) GPA	19	.24	21	11	18	33	-.13	-.03
9. HS GPA in Solids	24	.27	21	15	22	32	-.11	.01
10. HS Science GPA	24	.27	17	12	14	26	-.12	-.01
11. HS Chem GPA	11	.11	-.02	12	04	16	-.06	.02
12. 16 PF-Factor G	-.10	-.11	.02	-.20	-.21	-.13	.11	.08
13. 16 PF-Factor Q ₂	-.02	.03	.01	.02	.11	-.04	-.10	-.08
14. MMPI-Mf	.00	.11	-.16	.10	.22	-.04	-.10	.02
15. MMPI-Sc	.13	-.07	.00	.03	.09	.11	-.05	-.01
24. English	17	.17	.40	.25	.26	.27	-.09	-.09
25. Growth and Development	17	-.00	.40	.44	.33	.05	-.17	-.09
26. Microbiology	40	.40	-.00	.45	.43	.02	-.19	-.11
27. Nursing 2	25	.44	.45	-.00	.55	.03	-.33	-.29
28. Pharmacology 1B	26	.33	.43	.44	.55	.02	-.19	-.17
29. Sociology	27	.05	.02	.03	.02	-.02	.02	.03
30. Ward Performance (Rotation 1)	-.09	-.17	-.19	-.33	-.19	.02	-.02	.55
31. Ward Performance (Rotation 2)	-.09	-.09	-.11	-.29	-.17	.03	.55	-.00

^a All decimal points omitted from the correlation coefficients.

^b Coefficients of .17 and .23 are significant at the .05 and .01 levels, respectively.

and evaluated as follows:

1. The most valid predictor of success in the nursing program was the California Reading Test—Comprehension, although the correlations for the most part were modest or low.
2. The second and third most valid predictors were given by grade point average earned in high school academic subjects (variable nine) and overall high school grade point average (variable eight), although in selected subjects the California Mathematics Test—Reasoning was almost as valid as these two other predictors.
3. With the possible exception of the low validities for the two criterion variables of Physiology and Nursing 2 the measure of spatial visualization (variable seven) was not particularly predictive of success in any course within the program.
4. The two self-report inventories—the 16PF and the MMPI—yielded correlations of virtually no predictive value, as reflected by the fact that among 240 possible validity coefficients only 12 were significant at the .05 level for the total sample and only 24 were significant at the .05 level for the successful sample of 96 candidates. Only two scales from each of the two self-report inventories yielded two or more validity coefficients significant at the .05 level for the sample of successful students. These scales are the ones cited in Tables 1 and 2.
5. Evidence from Tables 1 and 2 points to the relatively low degree of intercorrelation among the criterion measures for both the total sample and the group of successful candidates. Although these coefficients might indicate that different characteristics were being evaluated in the several courses, their relatively low magnitude in relation to observations made by the writers in other school-oriented contexts would suggest that the reliability of the grading process might be open to question.
6. Among the intercorrelations of the criterion measures (as reported in Table 2) it should be noted that the correlations of the ward adjustment measures are near zero or negative to a slight degree, although in seven instances, significantly, with the measures in the other courses.

Conclusions. It is apparent from the relatively low magnitude of the correlations reported that despite the probable presence of some restriction in range of talent consideration needs to be given not only

to the introduction of other measures for the selection and placement of students in the nursing program but also to possible modifications in supervisory and in-service activities directed toward the improvement of the evaluation process. Although reliability data are not present for the criterion measures with the single exception of the two measures of ward performance which intercorrelated only .55, it would appear that efforts aimed at the more nearly precise definition of course objectives and the use of more reliable and valid methods for evaluation of performance could lead to substantially higher criterion-related validity coefficients.

REFERENCE

- Michael, W. B., Haney, R., and Jones, R. A. The predictive validities of selected aptitude and achievement measures and of three personality inventories in relation to nursing training criteria. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1966, 26, 1035-1040.

A NOTE ON THE VALIDITY OF TWO MEASURES OF HIGH SCHOOL RANK

GERALD W. McLAUGHLIN
United States Military Academy
West Point, New York

At present, two main ways exist for a college to utilize an applicant's high school academic performance. The first of these is the use of grades in high school, either an average or a weighted average of selected grades. In a recent report, the use of four weighted grades in multiple regression analyses resulted in median validities of .44 to .55 for a sample of 398 colleges (Munday, 1967).

The second method of incorporating this information is to use a measure of high school rank (HSR). The rank is usually expressed as a form of $(1 - R/N) \times 100$, where R is the applicant's rank in his graduating class and N is the size of the graduating class. It also has validity for predicting college academic performance (Borgatta and Bohrnstedt, 1969; Lavin, 1965).

Purpose. Both of these measures customarily improve the validity of predicted freshman grades. However, the use of HSR seems to be intuitively more appealing because of its simplicity. The purpose of this paper was to propose modification in the measurement of a college applicant's graduating rank in his high school class.

Procedure. The two measures of HSR used in this study were (a) the standard predictor (HSR1) computed as $(1 - R/N) \times 100$, and (b) a modified form (HSR2) computed as $(1 - R/CN) \times 100$ where CN is the number of those in the applicant's high school class planning to attend college. The size of the college bound segment was furnished by an official at the applicant's school. The sample consisted of 194 cadets entering the United States Military Academy in 1968 on whom complete data were available.

TABLE 1

Descriptive Statistics for Two Measures of High School Rank Scores

Rank Index*	\bar{X}	SD	SAT-V	SAT-M	ENG-C	MATH-A	GPA	HSR1
HSR1	84.79	13.74	.182	.157	.183	.234	.440	
HSR2	72.75	21.50	.237	.204	.252	.294	.549	.799

* The symbols HSR1 and HSR2 are described in the text.

The major criterion was a student's overall academic grade point average (GPA) at the end of his freshman year at the Academy. The degree of relationship between these rankings and the performance on each of four standardized tests of the College Entrance Examination Board (CEEB) was determined.

Findings. The results are shown in Table 1. The modification of the denominator in the High School Rank Score significantly improved the validity of the score for this sample ($p < .01$). This correction in rank is a logical one, since the HSR score is a measure of an applicant's performance relative to his peer group. If he plans to attend a college, a first approximation of his academic peer group size is the number of those in his graduating class also planning to attend college. This correction is also relative to the quality of high school, since the score of an applicant from a school sending everyone to college would be greater than the score of an individual with the same rank in an equivalent size graduating class where fewer graduates went to college.

Conclusion. The relationships of the measures with the CEEB tests suggest that HSR can provide a unique contribution to predicting an individual's academic grade point average.

REFERENCES

- Borgatta, E. F. and Bohrnstedt, G. W. The use of the Quick Number Test in the prediction of academic performance. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 921-925.
- Lavin, D. E. *The prediction of academic performance*. New York: Russell Sage Foundation, 1965, p. 52.
- Munday, L. Predicting college grades using ACT data. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 401-406.

MULTIVARIATE VALIDITY OF THE OTIS-LENNON MENTAL ABILITY TESTS PRIMARY I LEVEL

BRAD S. CHISSOM AND JERRY R. THOMAS

Georgia Southern College

MEASURES of academic achievement in the form of teacher ratings are often used as the criterion measure in validity studies. One of the limitations of teacher ratings is that they ignore the specificity of learning in the several areas comprising total achievement. A more complex rating based on specific areas of achievement would seem to offer greater possibilities for effective measurements. This study employed a complex teacher rating scale of kindergarten children as a criterion measure for validating the Otis-Lennon Mental Ability Test Primary I Level (1967). A validity coefficient was obtained through the use of canonical correlation by correlating the parts of the Otis-Lennon with the parts of the teacher ratings, a procedure suggested by Mukherjee (1966).

Studies employing teacher ratings of kindergarten children as measures of achievement have been conducted by Koppman and LaPray (1969) and Meyers, Attwell, and Orpet (1968) with some degree of success. For kindergarten children, a single teacher rating had only a moderate correlation with a single objective measure of academic aptitude.

Method—Subjects. The children employed as subjects in this study consisted of two classes of 20 kindergarten children ($N = 40$) from the Georgia Southern College Laboratory School. The mean age for the group was 67.48 months at the time of test administration (March, 1971).

Instruments. Two instruments were employed in this study. The first was a complex teacher rating scale in which the two kindergarten teachers were asked to assign a numerical rating from one

to nine for each child in four academic areas: (1) Reading (2) Quantitative, (3) Verbal, and (4) Listening. The second measure was the Otis-Lennon Mental Ability Test Primary I. The Otis-Lennon MAT consists of fifty-five items divided into two parts. Part I requires the subject to identify the different picture in a group of four pictures, and Part II directs the subject to select the correct picture corresponding to a verbal description. Both parts of the test are administered orally to the subject.

Results. Reliability for the total score on the teacher rating measure, which was calculated by Cronbach's Alpha, was estimated to be .92. The Otis-Lennon MAT reliability coefficient for total score, which was computed using the split-half odd-even method and then increased by the Spearman-Brown Formula, was equal to .91.

The resulting canonical correlation between the two sets of variables was .76, significant at less than the .001 level.

Examination of Table 1, which contains means, standard deviations, and beta weights, indicates differential weightings for the teacher ratings. Reading readiness and quantitative ability carry the heaviest weights, while listening ability is weighted negatively. The two parts of the Otis-Lennon are weighted approximately equal.

Summary. This study has demonstrated that teacher ratings appear to be more useful when they are composed of several parts. Further, canonical correlation analysis is a feasible technique for assessing validity when both measures incorporate part scores or subtests as frequently found in academic and intellectual measures.

TABLE 1

Means, Standard Deviations, Beta Weights, Reliabilities, and Canonical Correlation

Variables	Means	Standard Deviations	Z-Score Beta Weights	Reliability Estimates
Teacher Ratings				.92
Reading	5.58	1.62	.871	
Quantitative	5.93	1.53	.524	
Verbal	6.13	1.22	.226	
Listening	6.18	1.38	-.280	
Otis-Lennon				.91
Part I	13.45	4.58	.749	
Part II	14.97	5.05	.663	

$R_c = .76$, $\chi^2 = 38.86$ ($df = 8$), $p < .001$

It should be emphasized, however, that cross-validation efforts with new samples would be necessary before these results can be generalized.

REFERENCES

- Koppman, P. S. and LaPray, M. H. Teacher rating and pupil reading. *The Reading Teacher*, 1969, 22, 603-608.
- Meyers, C. E., Attwell, A. A., and Orpet, R. E. Prediction of fifth grade achievement from kindergarten test and rating data. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 457-463.
- Mukherjee, B. N. Application of canonical correlation analysis to learning data. *Psychological Bulletin*, 1966, 67, 9-21.
- Otis, A. S. and Lennon, R. T. *Otis-Lennon Mental Ability Tests*. New York: Harcourt, Brace, and World, 1967.

THE CONCURRENT VALIDITY OF THE SPRIGLE SCHOOL SCREENING READINESS TEST FOR A SAMPLE OF PRESCHOOL AND KINDERGARTEN CHILDREN

MARIA S. A. SEDA

Perris Union High School District, California

JOAN J. MICHAEL

California State College, Long Beach

Problem. It was the primary purpose of this study to determine the degree of relationship between the Sprigle School Readiness Screening Test (Sprigle) and the Metropolitan Readiness Test (MRT) with the view to substituting the Sprigle for the Metropolitan. Secondly, it was the purpose to investigate the relationship between scores on the Peabody Picture Vocabulary Test (both Peabody IQ score and Peabody raw score) and the MRT.

Procedure and subjects. One hundred children (25 preschool and 75 kindergarten in suburban Southern California), ranging in age from four years and 10 months to six years and nine months, were given the three tests. Starting in June 1970 and continuing through January 1971, the MRT (Form A), the Sprigle, and the Peabody (Form A) were administered in random order according to the manuals of directions. Regardless of the order, there was a week's separation between the administration of the MRT and the random administration of the Sprigle and the Peabody in light of the time involved in giving the Metropolitan and the limited attention span of this age child.

Results. As shown in Table 1, there was a correlation of .73 ($p < .01$) between the Sprigle and the MRT. Further the correlation between the Peabody IQ scores and the MRT scores was .58

TABLE 1

Correlation Matrix for the Sprigle School Screening Readiness Test, the Peabody Picture Vocabulary Test, and the Metropolitan Readiness Test

Variable	Variable			
	Sprigle raw scores	Peabody IQ scores	Peabody raw scores	Metropolitan raw scores
<i>Sprigle</i> raw scores55	.61	.73
<i>Peabody</i> IQ scores	.5590	.45
<i>Peabody</i> raw scores	.61	.9058
<i>Metropolitan</i> raw scores	.73	.45	.58	...

Note.—All correlation coefficients were significant beyond the .01 level.

($p < .01$); whereas the correlation between the Peabody raw scores and the MRT scores was .45 ($p < .01$).

It should also be noted that when a multiple R was computed between the composite made up of the Sprigle and Peabody IQ scores on the one hand and scores on the MRT on the other, a correlation of .73 was found. Similarly, when Peabody raw scores were substituted for the IQ scores, a multiple R of .75 was found. In neither case was there a statistically significant increment in R with the addition of the Peabody.

Discussion of Results. It might be of incidental interest to note that when the correlation of .73 between the Sprigle and the MRT was compared with the correlation of .58 between the Peabody raw score and the MRT, the Hotelling's t_{d_r} value (Guilford, 1965, p. 190) was 2.53 ($p < .05$). Further, when the correlation of .73 between the Sprigle and the MRT was compared with the correlation of .45 between the Peabody IQ scores and the MRT, the t_{d_r} value was 4.27 ($p < .01$).

Thus, in light of the high degree of correlation between the Sprigle and the widely used MRT (.73) and the fact that the addition of neither the Peabody IQ score nor raw score produced any significant increment in R , there seems to be evidence of concurrent validity of the Sprigle by itself as an effective screening device. This conclusion is further supported by the finding that the Sprigle can demonstrate greater potential predictive validity than can the

Peabody (either raw score or IQ score) when the criterion variable is raw score on the MRT.

Since the general long term goal of this study was to increase success and decrease failure in primary grades by identifying a test which would effectively predict school readiness, it was of interest to find not only an instrument which would be quick to give and valid, but also one which would provide the most information about the process a child uses as he attacks the tasks given to him. In addition to the high degree of correlation with the MRT, other salient features of the Sprigle are that it can be administered in 10 to 15 minutes by a nonprofessional and that diagnostic information pertaining to how each task is undertaken can be obtained by its individual administration.

It is, therefore, recommended that school districts investigate the use of the Sprigle as a possible quick screening instrument for kindergarten children which can yield information to assist teachers and administrators in assessing proper placement for entering school age children.

REFERENCE

- Guilford, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill Book Company, 1965.

A MULTITRAIT-MULTIMETHOD VALIDATION OF MEASURES OF STUDENT ATTITUDES TOWARD SCHOOL, TOWARD LEARNING, AND TOWARD TECHNOLOGY IN SIXTH GRADE CHILDREN

SOL M. ROSHAL, IRENE FRIEZE, AND JANET T. WOOD

Institute for Development of Educational Activities, Inc.

EDUCATORS and parents are becoming increasingly aware of the importance of student attitudes. The child is often expected not only to learn the required subject matter, but also to enjoy school and to look forward to learning new things. Also, there is concern among industrial leaders as well as educators that children appreciate the benefits of technology and that they not be afraid of the many machines in their environments. Previous studies (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, and York, 1966) have shown that student attitudes are related to school achievement. In their massive studies of United States schools, Coleman, et al. found that attitudes towards school and learning were significant indicators of verbal skills in sixth graders. Other studies have shown positive, but often nonsignificant relationships between grades and achievement scores and attitudes (Jackson and Lahaderne, 1967; Brodie, 1964).

Even with the increasing interest in measures for assessing student attitudes, there are few existing instruments in the literature. Many of those which do exist are parts of larger instruments. Often these subtests do not have tested reliability or validity as independent measures (Coleman, et al., 1966). Other instruments measure only attitudes towards school (Flanders, 1965; Jackson and Getzels, 1959) and often do not report validity data. Thus, there is a need for reliable and validated instruments to measure various important school attitudes. The present study involved the

validation of three student self-report attitude measures: Attitude Toward School (ATS), Attitude Toward Learning (ATL), and Attitude Toward Technology (ATT).

Validity assessment of attitude scales is difficult in general because absolute criteria for knowing who has a positive or negative attitude are not readily attainable. Usually, the only data available to validate an instrument are correlations with other measures of equally low reliability and/or validity. The most promising approach to this difficult problem of validity appraisal of attitude scales is the multitrait-multi-method matrix proposed by Campbell and Fiske (1959). Its use requires that several traits be assessed with several independent methods. In the Campbell and Fiske terminology the "traits" of this study were the three types of attitudes; the three "methods" used to assess the attitudes were the newly developed instruments, teacher ratings, and peer ratings.

Method—Source and selection of Items. Large numbers of items were constructed on the basis of content validity (items believed by educational specialists to measure the respective attitude) for each of the three scales. Attitude Toward School (ATS) items assessed feelings about school as an institution, about teachers and other school personnel, classmates, school subjects, and the classroom. Items for Attitude Toward Learning (ATL) were concerned with the student's general interest in the world, curiosity, interest in school subjects, reading, hobbies, and other learning activities. Attitude Toward Technology (ATT) had items from three general areas: personal control and understanding of machines, man's ability to control technology, and the positive benefits of technology. After several preliminary item analyses studies, which utilized factor analyses and item-total correlations, the final versions were constructed. They consisted of 25 items for ATS and ATL and 24 for ATT. Both positively and negatively worded items were used to control for response bias. Items were answered on 5-point Likert scales. Sample items from the final version of each scale are given in Table 1.¹

Administration procedures. The three scales, ATS, ATL, and ATT, were administered along with other questionnaires and

¹ Inquiries regarding the instruments and the development procedures may be sent to I[D]E[A], Research Division, 1100 Glendon Avenue, Suite 950, Los Angeles, California 90024.

TABLE 1
Sample Items

ATTITUDE TOWARD SCHOOL

10. $I \left\{ \begin{array}{l} \text{a. always} \\ \text{b. usually} \\ \text{c. sometimes} \\ \text{d. rarely} \\ \text{e. never} \end{array} \right\} \text{ hate school.}$

13. $\text{Teachers in this school are } \left\{ \begin{array}{l} \text{a. always} \\ \text{b. usually} \\ \text{c. sometimes} \\ \text{d. rarely} \\ \text{e. never} \end{array} \right\} \text{ friendly.}$

ATTITUDE TOWARD LEARNING

2. $\text{School subjects are } \left\{ \begin{array}{l} \text{a. always} \\ \text{b. usually} \\ \text{c. sometimes} \\ \text{d. rarely} \\ \text{e. never} \end{array} \right\} \text{ boring.}$

7. $\text{Whenever I go on a trip, I learn } \left\{ \begin{array}{l} \text{a. lots of} \\ \text{b. many} \\ \text{c. some} \\ \text{d. a few} \\ \text{e. no} \end{array} \right\} \text{ new things.}$

ATTITUDE TOWARD TECHNOLOGY

1. $I \left\{ \begin{array}{l} \text{a. strongly agree} \\ \text{b. agree} \\ \text{c. partly agree, partly disagree} \\ \text{d. disagree} \\ \text{e. strongly disagree} \end{array} \right\} \text{ that most new inventions help people live better.}$

11. $I \text{ could } \left\{ \begin{array}{l} \text{a. always} \\ \text{b. usually} \\ \text{c. sometimes} \\ \text{d. rarely} \\ \text{e. never} \end{array} \right\} \text{ learn how to fix almost anything.}$

peer ratings to a sample of 610 sixth grade students in 13 public schools. Their average Lorge Thorndike verbal IQ was 101.4 with a standard deviation of 15.7. There were approximately equal numbers of boys and girls. The sample ranged from lower middle to lower upper class in socioeconomic status (as judged by school district personnel).

For peer ratings students were asked to list two or more names, including themselves if they desired, for each of the following questions:

Which kids in this class really seem to enjoy school a lot?

Which kids are really interested in learning about a lot of different things?

Which kids in this class really enjoy using machines or fixing things that break down?

The peer ratings were then computed by counting the number of times each student was chosen, dividing by the number of students in the class (to equate the number of possible choices for different sized classrooms) and multiplying by 25 (to make the numbers whole values representing approximately the number of times chosen). Zero scores (i.e., no choices) were eliminated.

Teachers were also asked to rate each of their students with respect to:

1. *Attitude toward school*

Does he like school or not? Does he look forward to school and enjoy being there?

2. *Interest in learning both in and out of school*

Does he enjoy learning new things? Is he generally curious?

3. *Interest and understanding of machines*

Does he feel comfortable with various machines? Does he understand that machines cannot purposefully harm him?

These ratings were done on 5-point scales from "very high" to "very low."

Results—Descriptive data. Mean scores for the three attitude scales were 3.2 for ATS, 3.4 for ATL and 3.2 for ATT. The standard deviations were .67, .48, and .38, respectively. Since the possible scores ranged from one for very unfavorable to five for very favorable, the overall averages were all slightly more positive than the neutral point. There was most variance in ATS scores and least in ATT.

The scales had relatively high reliabilities as measured by Alpha coefficients (a form of K-R 20 or split-half reliabilities for multi-point responses) as shown in Table 2. ATS was most consistent with an Alpha of .93. ATL had a value of .84; and ATT, only .68.

Factor analyses (with orthogonal rotations) indicated that the ATS items had high loadings on one main factor—an outcome indicating the possible unidimensionality of this attitude. Three major factors emerged for the ATT corresponding to the three dimensions originally used for content validity. ATL was a more complex test that yielded four major orthogonal factors. These

were tentatively labeled as:

1. Boredom or lack of interest in life.
2. Enjoyment of learning new things.
3. Interest in school learning.
4. Interest in reading.

Multitrait-multimethod analysis. Analysis of the trait and method intercorrelations (shown in Table 2) by the Campbell and Fiske criteria indicated that all three instruments met most of the criteria.

The criteria and the relevant indices were:

1. The reliability of each trait measure should be significantly greater than zero and it should be greater than the correlation of the trait with the other trait measures. Reliabilities of the three scales (ATS, ATL and ATT) were well over the .09 needed for significance and were also greater than other scale intercorrelations (Matrix 1).

2. Diagonal values (italics) in validity matrices (Matrices 2, 3 and 5) should be greater than other values in the same row or column within that matrix. Thus, in Matrix 2, the validity correlation of teacher ratings of student school attitudes with ATS is .30. This is greater than the other values in the row (.27 and .01) and column (.18 and .00). Including the row and column comparisons for all three matrices, there are a total of 6 comparisons for each scale.

- a. ATS: of the 6 tests, 5 met this criterion
- b. ATL: 3 of 6 met criterion
- c. ATT: 4 of 6 met criterion

3. Correlations of traits between different methods should be higher than correlations of different traits within the same method. This criterion was not met since teacher ratings, peer ratings, and self-reports of the three traits tended to correlate more with each other within a given matrix than with each other across methods.

4. The same pattern of interrelationships should be found within all methods. This criterion was met for all instruments. ATS and ATL tended to correlate with each other and not with ATT. Correlations of ATT with ATL were higher than of ATT with ATS.

Discussion. All three scales yielded promising validation indications as measured by the multitrait-multimethod criteria. In

fact, a review by Campbell and Fiske (1959) indicated that few instruments in the literature at that time met more than one of two of these criteria while all three scales here met many of them. The reliabilities of the three instruments were also as high as or higher than similar instruments.

ATS, which measures the student's general attitude towards school as an institution, might be used by educators to measure feelings about school. It is probably relatively sensitive to attitude changes (although further studies of this are needed). As shown in Table 2, the correlation of ATS and ATL was .68 as might be expected because of the inclusion of school learning items on ATL. However, the multitrait-multimethod correlations do give support for the independence of the two instruments even though both teachers and peers had some difficulty differentiating the two concepts. ATL, which indicates a more general orientation toward learning, probably does reflect more of a personality trait than does ATS and thus would not be so susceptible to short term changes as are attitudes toward school.

ATT, which measures the student's perception of his ability to fix and run machines, his belief in man's general ability to control machines, and his beliefs about the positive benefits of technology, correlates very low with ATS and ATL. ATT might be of interest to educators with curricula related to the use of technology, or with projects which have machines (such as teaching machines) as part of the program. Also, realistic concepts of machines are important for daily living in a complex environment.

Any of the scales may be administered independently or in combination for elementary school assessment. They are presently being used with children in third through fifth grades as well as with children in the sixth grade. Although the reading difficulty of words used on the scales was purposefully kept low, use with average readers below the fifth grade is not recommended, however, unless the items are read aloud. Normative data for sixth grade pupils are available for all three scales.

REFERENCES

- Brodie, T. A. Attitude toward school and academic achievement. *Personnel and Guidance Journal*, 1964, 43, 375-378.
- Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York R. L. *Equality of Educational Opportunity*, Office of Education, United States Department of Health, Education, and Welfare, 1966.

Flanders, N. A. Teacher influence, pupil attitudes, and achievement. *Cooperative Research Monograph*, No. 12. (OE-25040) Washington D.C.: United States Government Printing Office, 1965.

Jackson, P. W. and Getzels, J. W. Psychological health and classroom functioning: a study of dissatisfaction with school among adolescents. *Journal of Educational Psychology*, 1959, 50, 295-300.

Jackson, P. W. and Lahaderne, H. M. Scholastic success and attitude toward school in a population of sixth graders. *Journal of Educational Psychology*, 1967, 58, 15-18.

DOGMATISM AND CONSERVATISM: AN EMPIRICAL FOLLOW-UP OF ROKEACH'S FINDINGS

FRANK COSTIN

University of Illinois at Urbana-Champaign

In partial support of his claim that the Dogmatism Scale measures "general authoritarianism," and is "relatively free of political content," Rokeach (1960, pp. 121-122) has cited low correlations between scores on his scale and scores on the Politico-Economic Conservatism Scale (Adorno, Frenkel-Brunswick, Levinson, and Sanford, 1950). However, since these correlations were consistently positive, Rokeach also concluded: . . . "The chances are somewhat better than even that a closed-minded person will be conservative rather than liberal in his politics." (p. 122).

Unfortunately, the measure of "conservatism" which Rokeach used consisted of only five items. In describing the construction of this instrument, Daniel Levinson pointed out that for practical reasons it became necessary to reduce it considerably from its original length, and readily conceded that the five items were "not enough to obtain an adequate measure of reliability, and hardly enough to be called a 'scale.'" (Adorno, et al., 1950, p. 168). Furthermore, the items tended to be broadly stated: e.g. "In general, full economic security is bad; most men wouldn't work if they didn't need the money for eating and living." "America may not be perfect, but the American Way has brought us about as close as human beings can get to a perfect society." (Adorno, et al., 1950, p. 169).

What relationship between dogmatism and conservatism might be obtained if (a) a more reliable measure of "conservatism" were used, (b) the items dealt with more specific and current issues, and (c) "conservatism" was more operationally defined? The na-

tional election campaign of 1970 provided an excellent opportunity to answer this question.

Method. A senatorial candidate from the Midwest set forth his position as a "conservative" in a newspaper advertisement by inviting readers to take a "test" to see how "conservative" or "liberal" they were. The "test" consisted of a series of 15 paired statements; in each pair one of the statements represented a "conservative" opinion, and the other a "liberal" opinion. For example:

I am for stronger laws to curb the sale and distribution of pornographic materials.

We should not restrict the freedom of publishers or movie producers to produce or sell any material they choose.

Law enforcement officials are too hard on protesters.

I am in favor of taking a stronger stand against protesters who engage in violence on the campus and in our communities.

In addition to the issues reflected in the above examples, other paired statements dealt with President Nixon, Vice-President Agnew, Vietnam, bussing to achieve racial balance in schools, Judge Hoffman and the Chicago 7, and the Federal welfare programs. Readers were asked to score their answers according to a key indicating which statements were "conservative" positions and which were "liberal" ones; thus, according to the advertisement, readers could compare their opinions with those of the candidate. (He supported all of the "conservative" statements).

Ten pairs of statements were selected from the advertisement to represent a minimum of redundancy and a maximum of specificity. Each pair included a "conservative" opinion and a "liberal" opinion on the same issue. The 20 statements were combined with the 40 items of the Dogmatism Scale, Form E (Rokeach, 1960, pp. 73-80), and all arranged in a random order to form a 60-item questionnaire labeled "Opinion Survey: Social-Psychological."¹ (All items on the Rokeach Scale are stated in the "dogmatic" direction).

During the summer of 1970 the "Survey" was administered to 78 students (40 men, 38 women), selected randomly from the subject pool of an introductory psychology course. (Eighty-five students had been asked to participate, but seven failed to re-

¹ A copy of this instrument may be obtained by writing to: Frank Costin, 731 Psychology Building, University of Illinois, Champaign, Illinois 61820.

port). Directions for responding to the items followed the standard instructions of the Dogmatism Scale (Rokeach, 1960, pp. 72-73); however, instead of scoring responses on a scale from +3 ("agree very much") to -3 ("disagree very much"), a scale from 6 to 1 was used, with 6 corresponding to +3 and 1 corresponding to -3. All dogmatism items and the 10 "conservative" statements were scored accordingly. Direction of scoring was reversed for the 10 "liberal" statements, so that "disagree very much" was assigned a score of 6, and "agree very much" a score of 1.

For the purpose of computing reliability coefficients (KR-20), responses were also scored as 1 or 0; in the case of a dogmatism item or a "conservative" statement, any degree of agreement was assigned a score of 1; for a "liberal" statement, any degree of disagreement was assigned a score of 1. Based on the responses of all 78 students, KR-20 was .78 for the Dogmatism Scale and .79 for the remaining 20 items. (Correlations between total scores obtained under dichotomy procedures and those obtained under the six-point system were .87 for the Dogmatism Scale and .90 for the other 20 items).

Results. Table 1 shows the correlations (*r*'s) between scores on the Dogmatism Scale and the scores on the 20 items measuring "conservatism"; it also reports the means of these scores. For all 78 students, the correlation between "dogmatism" and "conservatism" was .56. This coefficient is significantly greater than any of those which Rokeach obtained when he correlated college stu-

TABLE 1
Relationship between Dogmatism and Conservatism

		Dogmatism ^a		Conservatism ^b		Dogmatism vs Conservatism (<i>r</i>)
		Mean	SD	M	SD	
Men	(<i>N</i> = 40)	131.4	28.2	51.2	13.5	.63*
Women	(<i>N</i> = 38)	129.0	17.9	47.1	11.7	.45*
Total	(<i>N</i> = 78)	130.3	23.6	49.2	12.8	.56*

Note—All respondents were students in an introductory psychology course.

^a Dogmatism was measured with the Rokeach Scale, Form E. Each of the 40 items was scored on a scale from 6 ("agree very much") to 1 ("disagree very much"); the higher the score the greater the dogmatism.

^b Items were 20 statements taken from a political advertisement by a self-announced "conservative" candidate for national office. Ten items were in a "conservative" direction, and scored on a scale from 6 ("agree very much") to 1 ("disagree very much"); the other ten items were in a "liberal" direction, and scored in reverse fashion; the higher the score, the greater the "conservatism."

* $p < .01$.

dents' scores on the Dogmatism Scale with their scores on the five-item Politico-Economic Conservatism Scale. The r 's he reported, and the significance level of each difference, were as follows: .11 (New York Colleges, $N = 207$, $p < .01$); .13 (Michigan State University, $N = 202$, $p < .01$); .20 (Michigan State University, $N = 153$, $p < .01$); .28 (Michigan State University, $N = 186$, $p < .05$). (Rokeach, 1960, p. 122).

Table 1 also shows that the correlation between "dogmatism" and "conservatism" was higher for men than for women (.63 vs .45); however, the p value for the difference between these two r 's was greater than .05.

These results indicate that the relationship between "conservatives" (political-economic-social) and Rokeach's interpretation of "closed-mindedness" may be stronger than he realized. The findings may also reflect the advantage of specifying operationally the "conservatism" one intends to measure. Of course, more extensive investigations of this kind need to be carried out to see whether such conclusions can be further supported, by using either the "conservative" and "liberal" statements employed in the present study (the issues are still lively ones) or some similar measure.

REFERENCES

- Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., and Sanford, R. N. *The authoritarian personality*. New York: Harper and Row, 1950.
- Rokeach, M. *The open and closed mind*. New York: Basic Books, 1960.

SPECIFIC ANXIETY THEORY AND THE MANDLER-SARASON TEST ANXIETY QUESTIONNAIRE

FRANK B. W. HARPER

University of Western Ontario, Canada

THE specificity theory of anxiety advanced by Mandler, Sarason and their colleagues in a series of publications since the early nineteen fifties (particularly Mandler and Sarason, 1952; Sarason, Mandler and Craighill, 1952; Sarason, Davidson, Lighthall, Waite, and Ruebush, 1960) hypothesizes that anxiety is attached to and aroused by specific situations and that it is more valid to measure anxiety by items pertinent to particular situations than by items which purport to measure anxiety in some general way. In line with this theory the Test Anxiety Questionnaire (TAQ) was developed to measure anxiety aroused by evaluative or testing situations (Mandler and Sarason 1952).

The TAQ has undergone a number of modifications, both of content and scoring, since its first inception. The most commonly used version is one which contains three sections: the first dealing with anxiety about group intelligence tests, the second dealing with anxiety about individual intelligence tests, and the third dealing with anxiety about course examinations. If the specificity theory were to be followed through to its logical conclusion, one would expect that the three sections would each be scored separately and that then these scores would be compared with the appropriate criterion variable; i.e., the score on anxiety about group intelligence tests would be compared with the results of group intelligence tests, and the score on anxiety about course examinations would be compared with examination results. Historically, however, this seems not to have been the scoring system used, and instead a total

score on all three sections has been employed as the anxiety variable. The use of such a total composite score can be criticized on the same grounds that are used to distinguish specific anxiety, e.g., test anxiety from general anxiety. If anxiety is really specific to situations, then one would expect that the most appropriate items for predicting the effects of anxiety on marks in achievement examinations would be items which deal specifically with responses to taking course examinations, and a similar argument could be advanced for anxiety about intelligence tests. Theoretically, a score obtained by adding together the scores on different kinds of anxieties, even if they are all classed generally as test anxieties, should not be so efficient in predicting a given criterion, as one which is specific to the criterion situation.

Purpose. The present study was concerned with comparing the concurrent validity of each of the three sections of the TAQ, plus its composite total, against the criterion of cumulative grade point average (CGPA) in college academic courses. The use of CGPA was intended to capitalize on a broader sampling of the course-examination anxiety domain, as compared with, say, a single semester grade on a particular course, in which there could be strong effects of chance and biased sampling.

The following hypothesis was tested: Of the four scores possible on the TAQ, taking each section by itself, and the total score of all three, the score on the section dealing with Anxiety about Course Examinations would correlate most highly—*negatively*—with CGPA.

Method—Subjects. Two samples of college students were studied. The first sample was 57 males and 168 females in the third year of their undergraduate career at the University of Minnesota. The second sample was 44 males and 43 females at the University of Western Ontario, Canada, all of whom had just been graduated with a Bachelor's degree.

Procedure. Each student was given the TAQ to complete in a regular class meeting, with instructions which advised him that the questionnaire was part of a research project on test anxiety. The college transcript from each student was obtained, and the cumulative grade point average calculated.

Scoring. There are a number of scoring systems for the TAQ. In this study each item was divided into five sections, and the student

marked which of the five he felt applied to him. The score for each section was the sum of the ratings for the items in that section. The total score was the arithmetic sum of the three sections.

The product moment correlations between the sections and the cumulative grade point averages are shown in Table 1, together with the multiple correlation coefficient (R) using each subsection in a three variable predictor regression equation.

In all four groups the highest correlations occurred, according to the hypothesis. Anxiety about Course Examinations correlated most highly negatively with Grade Point Average. All these correlations between CGPA and Course Examination Anxiety were significantly different from zero, a fact which was not true of the Total Score Correlations. Only one of the latter types of correlation reached significance, that of the Minnesota male sample. The correlations for the sections on Anxiety about Groups and Individual Testing were with one exception not significant. Calculation of the multiple correlation coefficient (R) predicting CGPA from the three subsections treated as independent variables, showed that R was numerically greater than the correlation for total score in each instance. Of course the contribution of Course Examination Anxiety to the multiple R was very high.

Discussion. The hypothesis that Anxiety about Course Examin-

TABLE 1

Product Moment Correlations of Cumulative Grade Point Averages (CPGA) with Test Anxiety Questionnaire (TAQ) Total and Subscale Scores (Decimal Points Omitted)

Section	Males		Females	
	Minn. ^a	Can. ^b	Minn. ^a	Can. ^b
Anxiety about Group IQ Tests	-.24*	-.12	-.12	-.13
Anxiety about Individual IQ Tests	-.11	-.07	.00	-.08
Anxiety about Course Exams	-.47***	-.29*	-.18**	-.25*
Multiple R^c	.49**	.30	.20	.26
Total Score	-.37**	-.17	-.11	-.19
N	57	44	168	43

^a Minnesota.

^b Canadian.

^c Based on treating the three sections of the TAQ as independent variables.

* $p < .05$.

** $p < .01$.

*** $p < .005$.

ations would correlate most highly, *negatively*, with cumulative grade point average was sustained by the results. The specificity theory of test anxiety was given further credibility by the findings. It appeared that even within the test anxiety domain, there were different kinds of test anxieties and that a test for one kind did not necessarily correlate with a test for another kind. An important consequence of the results is that researchers who have been content to use only the Total Score of the Test Anxiety Questionnaire as their measure of test anxiety might possibly have reached erroneous conclusions about the validity of the scale. If the criterion measure, e.g. intelligence test performance, is correlated with the appropriate section rather than with the total score, significant results might be obtained. Alternatively a multiple correlation using the three subsections as independent variables in a regression equation would be preferable to using the total score. Sassenrath's pessimistic comments (Sassenrath, 1967) on the validity of the Test Anxiety Questionnaire might be re-examined in the light of these suggestions.

REFERENCES

- Mandler, G. and Sarason, S. B. A study of anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, 47, 166-173.
- Sarason, S. B., Mandler, G., and Craighill, P. G. The effects of differential instructions on anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, 47, 561-565.
- Sarason, S. B., Davidson, K., Lighthall, F., Waite, R. R., and Ruebush, B. K. *Anxiety in Elementary School Children*. John Wiley and Sons, New York, 1960.
- Sassenrath, J. M. Anxiety, aptitude, attitude and achievement. *Psychology in the Schools*, 1967, 4, 341-346.

HOSTILITY AND LEARNING: A FOLLOW-UP NOTE

FRANK COSTIN

University of Illinois at Urbana-Champaign

A recent study (Costin, 1970) found significant negative correlations between examination scores of 50 male students in an introductory psychology course at the University of Illinois, and their precourse hostility scores, as measured by the Scrambled Sentence Test (Costin, 1969). Without necessarily implying a direct cause and effect relationship, the investigator interpreted the role of hostility in this context as an "interference" with learning. Additional data gathered independently of this study were consistent with such an inference: end-of-semester grade point averages of male students enrolled in the Special Educational Opportunity Program at the University ($N = 129$) were found to be negatively correlated with presemester scores on the Scrambled Sentence Test.

The purpose of the present study was to discover whether the Scrambled Sentence Test might also be a negative predictor of achievement in a very different kind of educational setting—a highly technical course at a military installation. If so, the inference that hostility may interfere with learning would gain greater plausibility.

Method. The subjects were 60 enlisted men at an Air Force Technical Training Center. They were enrolled in a 16 weeks course dealing with principles of meteorology and their application to observing and recording weather phenomena. Teaching-learning activities included lectures, demonstrations, laboratory work, discussion and assigned readings. Course achievement was evaluated with objective examinations and practical performance tests.

The Scrambled Sentence Test was administered to all students at the beginning of the course. Prior to entering the course they had

completed the Air Force Qualification Test, a group measure of general mental ability.

The Scrambled Sentence Test was the only measure the investigator was permitted to introduce into the classroom procedures; furthermore, no pretest of knowledge or skills concerning weather observation was being used by the instructional staff at this time. Thus, whatever advantages or disadvantages such precourse information might reflect could not be assessed. However, the investigator did not consider the lack of such information as crucial for demonstrating additional evidence concerning the relationship between hostility and course achievement, since one might reasonably assume that individual differences in knowledge of weather observation principles and techniques were probably minimal at the beginning of the course. It was also assumed that failure to control for such precourse knowledge and skill need not necessarily preclude interpreting end-of-course achievement as "learning," even though conventional definitions of learning usually incorporate the concept of "change." As Bereiter (1963) has observed, "many of the situations in which people change are not situations in which change is a meaningful variable. Situations involving uniform training procedures aimed at bringing subjects to a certain terminal level of performance are of this type . . . [pp. 13-14]" (Bereiter used an electronics course to illustrate the point.)

Results. Table 1 shows the intercorrelations of scores for hostility, course achievement, and mental ability. As the table indicates, the zero-order correlation between hostility and achievement was $-.41$ ($p < .01$); the partial correlation (ability held constant)

TABLE 1

Intercorrelations of Male Air Force Students' Scores on Hostility Test, Scores on the Air Force Qualification Test (AFQT), and Total Course Achievement. (N = 50)

	Zero-order r Achievement	Ability	Partial r with achievement, ability held constant
Hostility	$-.41^*$	$-.18$	$-.39^*$
AFQT	$.61^*$		

Note.—Hostility was measured with the Scrambled Sentence Test, Form C (Costin, 1969); the higher the score, the greater the hostility. (Maximum possible score was 30; mean = 10.9, SD = 4.2). Achievement was based on total number of points accumulated on written and performance tests. (Maximum possible score was 300; mean = 251.1, SD = 12.9.)

* $p < .01$, two-tailed.

was $-.39$ ($p < .01$). This finding is consistent with that obtained for the 50 male students in the previous study of introductory psychology; in that instance, with ability held constant, the partial correlations between Scrambled Sentence Test scores and achievement scores on two objective examinations of principles and other empirical generalizations were $-.41$ and $-.45$ ($p < .01$). (When both ability and precourse knowledge were controlled, the r 's were $-.40$ and $-.44$ respectively.)

The fact that the Scrambled Sentence Test was a negative predictor of course achievement in these two different teaching-learning situations lends further support to the possibility that hostility tends to interfere with learning. For the present this inference should be restricted to men, since the correlations between achievement and hostility scores for women in the psychology course ($N = 51$), while negative, were not significant ($p > .05$). Although the data of the psychology study suggest that hostility may be less of an interference for women than for men, such a conclusion must be held in abeyance, since tests of significance for differences between the r 's yielded probability values greater than .05. Investigations are now under way to discover whether negative relationships between hostility and learning may indeed be greater for men than for women.

REFERENCES

- Bereiter, C. Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.) *Problems in measuring change*. Madison: University of Wisconsin Press, 1963, 3-20.
- Costin, F. The Scrambled Sentence Test: A group test of hostility. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 461-468.
- Costin, F. Hostility and learning in an introductory psychology course. *Psychology in the Schools*, 1970, 7, 370-374.

PREDICTING THE BEHAVIOR OF INSTITUTIONALIZED DELINQUENTS WITH—AND WITHOUT—CATTELL'S HSPQ¹

VERNON O. TYLER, JR.
Western Washington State College

ROBERT F. KELLY
Spokane, Washington

THE High School Personality Questionnaire (HSPQ) is one of a family of tests developed by Cattell and his co-workers (Cattell, Eloff, and Coan, 1958) to measure 14 personality factors of children ages 12-17.

There has been considerable research on the factorial validity of these tests (e.g., Cattell, 1957; Cattell and Scheier, 1961) and efforts have been made to match their personality factors to rating data (e.g., Becker, 1960; and Schaie, 1962). However, Vernon (1965) stressed the need for validation against a variety of external criteria. There have been a few normative studies on delinquent populations (Pierson and Kelly, 1963a, b; Stern and Grosz, 1969). The HSPQ has improved the prediction of academic success (e.g., Butcher and Gorsuch, 1960) and selected delinquents from general population.² However, this test has seldom predicted be-

¹ Data were collected when the authors were at the Fort Worden Diagnostic and Treatment Center, Port Townsend, Washington. Fort Worden is an institution of the Washington State Department of Institutions, Division of Juvenile Rehabilitation. Many thanks are due Gus Lindquist, Superintendent, and Assistant Superintendents Robert Tropp and Robert Koschnick for their support and encouragement of this research. Appreciation is also due the many cottage staff who gave their time to complete the diagnostic ratings. Portions of this paper were read at Western Psychological Association, San Diego, March, 1968.

² I. H. Scheier, personal communication, January 11, 1962.

havior *within* a delinquent population. Pierson (1964) and Pierson, Cattell, and Pierce (1966) showed the HSPQ and other Cattell tests are sensitive to personality and academic changes in institutionalized delinquents. Tyler and Kelly (1962) found that HSPQ's administered in a diagnostic center for court-committed delinquent boys predicted ratings of behavior of these youths in treatment institutions several months later with specification-equation "multiple R 's," (Cattell, Beloff, and Coan, 1958) ranging from .18-.55.

The present study investigated the efficiency of the HSPQ and diagnostic ratings in predicting inmate characteristics.

Procedure. Subjects were 168 male offenders ages 14-18 housed in a state diagnostic center. Forms A and B of the HSPQ (1958 edition) were administered to these boys and they were rated by cottage staff who knew them for several weeks on 16 dimensions covering such behaviors as table manners, group functioning, work habits, hostility to staff, and masculinity (See Table 1). Because of staff and inmate turnover and scheduling, the raters were not trained, and the ratings were not made on all boys on one scale at a time, nor by the same number of raters; consequently, the reliability of these ratings probably was attenuated to an unknown degree. Several months later at four forestry camps, the boys' camp counselors³ rated them on these 16 dimensions and nine others, 25 criterion dimensions in all (See Table 1). These 25 rating scales were originally developed around characteristics the camp counselors considered most often in describing inmate behavior. Efforts were made to word the scales in the counselors' own language. At the time of administration, the counselors knew the boys rated quite well. Their only instructions were to attempt to use all categories on the scale in a pattern at least resembling a normal distribution. Reliabilities of tests (using raw scores) and ratings were calculated⁴ (Ebel's intraclass correlation; Guilford, 1954) and multiple re-

³ The authors wish to thank Tom Girard, Judson Turner, Richard Vernon, and Richard Philpott, superintendents, respectively of Cedar Creek, Mission Creek, Capitol Forest and Spruce Canyon youth camps and their camp counselors who generously gave their time and effort to rate the behavior of their charges.

⁴ Many thanks are due Dr. David B. Dekker, Director of the Research Computer Laboratory, University of Washington, for his guidance in writing the reliability program and processing data.

TABLE 1
Diagnostic Center and Forestry Camp Rating Scales

Scale No.		
1. Excellent table manners.	1 2 3 4 5 6 7 8 9	Disgusting table manners.
2. Functions well in a group (3 or more persons).	1 2 3 4 5 6 7 8 9	Does not function well in any group.
3. Not accident prone.	1 2 3 4 5 6 7 8 9	Very accident prone.
4. Always tells the truth. ^a	1 2 3 4 5 6 7 8 9	A regular liar.
5. Feels guilty when does something wrong.	1 2 3 4 5 6 7 8 9	Does not feel guilty when does something wrong.
6. Always does a good day's work.	1 2 3 4 5 6 7 8 9	Never does a good day's work.
7. Not hostile to staff.	1 2 3 4 5 6 7 8 9	Extremely hostile to staff.
8. Does not usually foul himself up.	1 2 3 4 5 6 7 8 9	Seems determined to mess himself up.
9. Pretty open and aboveboard.	1 2 3 4 5 6 7 8 9	Can't trust him out of my sight; very sneaky, always up to something.
10. Satisfactory adjustment on work crew. ^a	1 2 3 4 5 6 7 8 9	Unsatisfactory adjustment on work crew.
11. Never have to be firm with him to keep him in line.	1 2 3 4 5 6 7 8 9	Got to be really tough with him to keep him in line.
12. An adequate placement (or foster home) exists for him when paroled. ^a	1 2 3 4 5 6 7 8 9	No adequate parole placement exists at this time.
13. A very likeable kid. ^a	1 2 3 4 5 6 7 8 9	A very unlikeable kid.
14. Calm	1 2 3 4 5 6 7 8 9	Very nervous.
15. Good parole bet; will make it.	1 2 3 4 5 6 7 8 9	Very poor parole risk; won't make it.
16. Very manly—can stand on his own two feet.	1 2 3 4 5 6 7 8 9	A real "mama's boy."
17. Attains long term goals (several months ahead).	1 2 3 4 5 6 7 8 9	Can't even attain a short term goal (one or two days).
18. A nice looking kid. ^a	1 2 3 4 5 6 7 8 9	Ugly kid.
19. Trusts staff.	1 2 3 4 5 6 7 8 9	Does not trust staff at all.
20. Never hits or pushes.	1 2 3 4 5 6 7 8 9	Often hits and pushes.
21. Talks freely in counselling sessions. ^a	1 2 3 4 5 6 7 8 9	"Clams up" in counselling sessions.
22. Seldom picked on by other boys.	1 2 3 4 5 6 7 8 9	Frequently picked on by other boys.
23. Thinking seems O.K. ^a	1 2 3 4 5 6 7 8 9	Thinking seems pretty crazy.
24. Satisfactory adjustment in camp (not on work crew). ^a	1 2 3 4 5 6 7 8 9	Unsatisfactory adjustment in camp (not on work crew).
25. Sexually normal. ^a	1 2 3 4 5 6 7 8 9	Abnormal sexual behavior.

^a Indicates scales used in forestry camps only.

gression equations set up⁵ for prediction of scores on each of the 25 camp rating scales using the Wherry-Doolittle test selection method with the Wherry shrinkage formula (Garrett, 1958).

Results. The equivalent forms reliability coefficients for the HSPQ with the Spearman-Brown correction for length ranged from .33 to .68 for the 14 test factors.⁶

On the 16 diagnostic center rating scales, some boys were rated by as few as three raters and others by as many as eight. Problems with this rather untidy data have prevented the computation of reliability coefficients; it is estimated that they would range from .50 to .80. Reliability of the mean criterion ratings ranged from .50 to .92 with a median of .86.

Multiple R 's corrected for shrinkage were calculated for predicting camp ratings with the HSPQ, with diagnostic center ratings and with a combination of HSPQ and diagnostic center ratings. The HSPQ alone predicted with R 's ranging from .17 to .41 with a median of .30, with all but one R significantly greater than zero ($p < .01$). Diagnostic center ratings estimated camp ratings with R 's ranging from .29 to .55 (median $R = .49$) with all R 's significant ($p < .01$). Combined HSPQ and Diagnostic Center ratings produced R 's from .34 to .61 (median $R = .53$) with all R 's significant ($p < .01$).

The data show a clear-cut trend: while the 25 mean camp ratings were predicted by the HSPQ alone with fairly sizeable R 's (median $R = .30$), the validities of the diagnostic center ratings were considerably higher (median $R = .49$); and the combination of HSPQ and diagnostic center ratings was the highest of all (median $R = .53$).

Discussion. The obtained equivalence reliability coefficients for the HSPQ factors were comparable to those reported by Cattell et al. (Cattell, Beloff and Coan, 1958). The mean camp criterion ratings showed substantial reliability as was the case in a previous study (Tyler and Kelly, 1962).

With one exception, all 75 R 's computed were highly significant ($p < .01$). More interesting, however, was the clear predictive

⁵ Acknowledgement and appreciation are due Gardner Rowley of the Mathematics Department and the Computer Center, Western Washington State College, for calculating the multiple regression equations.

⁶ Complete data for the study may be obtained from the senior author.

superiority of the diagnostic center ratings over the HSPQ. It appears that for the practical task of predicting treatment institution behavior, diagnostic rating shows more promise than diagnostic testing. The slight gain in predictive power produced by combining tests and ratings hardly seems worth the cost of testing.

At best though, the predictive efficiency of any of these procedures is low. With R 's in the low .50's, only about 25 per cent of the variance in the camp ratings is accounted for. However, with further work, the predictive power of the diagnostic center ratings could probably be improved to the point of genuine diagnostic usefulness. With inmates remaining in diagnostic cottages for short stays of six weeks or less, the problems of rating are difficult. However, with trained raters and the rating of all boys in the cottage on one scale at a time (as was done in the forestry camps), improved predictions should result.

Of course, it must be noted that since for each R , a small number of predictor variables (2-8) was selected from a large pool of variables (14-30) the danger of shrinkage was present.

Cross validation is needed, but even more important would be a replication of the study with improved diagnostic ratings.

REFERENCES

- Becker, W. C. The matching of behavior rating and questionnaire factors. *Psychological Bulletin*, 1960, 57, 201-212.
- Butcher, J. and Gorsuch, R. Predicting Academic Achievement in Junior High School and High School. In *IPAT Information Bulletin #4*. Champaign, Illinois: Institute for Personality and Ability Testing, 1960.
- Cattell, R. B. *Personality and motivation structure and measurement*. New York: World Book, 1957.
- Cattell, R. B., Beloff, H., and Coan, R. W. *Handbook for the IPAT High School Personality Questionnaire*. Champaign, Illinois: Institute for Personality and Ability Testing, 1958.
- Cattell, R. B. and Scheier, I. A. *The meaning and measurement of neuroticism and anxiety*. New York: Ronald Press, 1961.
- Garrett, H. E. *Statistics in psychology and education*. (5th ed.) New York: Longmans Green, 1958.
- Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.
- Pierson, G. R. Current research in juvenile delinquency with IPAT factored instruments. *IPAT Information Bulletin #11*. Champaign, Illinois. Institute for Personality and Ability Testing, 1964.
- Pierson, G. R., Cattell, R. B., and Pierce, J. A demonstration by

the HSPQ of the nature of the personality changes produced by institutionalization of delinquents. *Journal of Social Psychology*, 1966, 70, 229-239.

Pierson, G. R. and Kelly, R. F. HSPQ norms on a state-wide delinquent population. *Journal of Psychology*, 1963, 56, 185-192. (a)

Pierson, G. R. and Kelly, R. F. Anxiety, extroversion and personality idiosyncrasy in delinquency. *Journal of Psychology*, 1963, 56, 441-445. (b)

Schaie, K. W. On the equivalence of questionnaire and rating data. *Psychological Reports*, 1962, 10, 521-522.

Stern, H. and Grosz, H. J. HSPQ personality measurement of institutionalized delinquent girls and their temporal stability. *Journal of Clinical Psychology*, 1969, 25, 289-292.

Tyler, V. O., Jr. and Kelly, R. F. Cattell's HSPQ as a predictor of the behavior of institutionalized delinquents. *Psychology Research Report No. 2*, November, 1962. Fort Worden Diagnostic and Treatment Center, Port Townsend, Washington.

Vernon, P. E. The HSPQ. In O. K. Buros. (Ed.) *Sixth mental measurement yearbook*. Highland Park, N.J.: Gryphon, 1965.

INTEREST PROFILES OF CLERGYMEN AS INDICATED BY THE VOCATIONAL PREFERENCE INVENTORY

DAVID L. SCHULDT

Wesley Foundation, Iowa City

ROBERT F. STAHMANN

The University of Iowa

THE literature related to the interests and personality patterns of clergymen does not include a report of the use of the Vocational Preference Inventory (VPI) with pastors (Schuldt, 1970). John Holland, the developer of the VPI, has reported an interest pattern for clergymen on his instrument (Holland, 1969). The purpose of the present study was to validate Holland's reported pattern for clergymen on the VPI with a sample of active United Methodist pastors.

Instrument and sample. Holland has proposed that the choice of a vocation is an expression of personality. The VPI consists of eleven scales which measure vocational interests and aspects of personality. The first six scales (Realistic, Intellectual, Social, Conventional, Enterprising and Artistic) measure specific interests and relate them to occupational environments. The remaining five scales (Self-Control, Masculinity, Status, Infrequency and Acquiescence) yield information about other aspects of the subject's personality (Holland, 1965).

The clergymen used in the study were randomly selected from among United Methodist clergymen serving Iowa churches as pastors. For the purpose of the study only those men who had served fewer than 15 years in the pastorate since their ordination were selected. The mean age of the respondents was 39.4 years. Seventy-three percent ($N = 55$) of the pastors sampled ($N = 75$) returned the

questionnaire materials which were administered by mail. A control sample of 105 employed adults, most of whom were college graduates, was drawn from the VPI Manual (Holland, 1965).

Results. Directions for VPI interpretation indicate that the highest score represents a dominant personality type and the four highest scores form a personality and interest pattern. Holland has reported a hierarchical pattern for ministers on the VPI of SAIE (Social, Artistic, Intellectual, Enterprising) (Holland, 1969). The VPI interest pattern for the clergymen sample tested in this study was also SAIE, which supported Holland's prediction. Table 1 summarizes the data for two groups: pastors and employed adult males. As can be observed from Table 1, the pattern for pastors is SAIE while that for employed adults is EICS (Enterprising, Intellectual, Conventional, Social).

The pastors' highest score on the VPI was on the Social scale; thus, they were of the Social personality type. The VPI Manual interprets high scores for males on this scale to indicate persons who are responsible, accepting of feminine impulses and roles, and are "facile and insightful in interpersonal relationships." Such persons have the ability to form "close" as opposed to "superficial"

TABLE 1
Differences between Vocational Preference Inventory Scores for Samples of United Methodist Pastors and Employed Adults

	Pastors SAIE (N = 55)		Employed Adults ^a EICS (N = 105)		t-test
	Mean	SD	Mean	SD	
1. Realistic	2.91	2.84	4.3	3.3	- 2.65*
2. Intellectual	4.20	4.23	7.0	4.6	- 3.76**
3. Social	8.13	4.00	5.4	4.0	4.1**
4. Conventional	2.25	3.15	4.4	3.5	- 3.81**
5. Enterprising	3.60	3.36	8.1	3.1	- 8.47**
6. Artistic	5.27	4.02	4.5	3.8	1.93
7. Self-Control	10.38	3.85	9.3	3.2	1.89
8. Masculinity	6.58	2.09	9.1	2.1	- 12.46**
9. Status	8.07	2.96	9.8	2.6	- 3.80**
10. Infrequency	6.64	3.27	4.0	2.5	5.69**
11. Acquiescence	10.22	5.15	13.4	5.4	- 3.59**

^a Sample drawn from the VPI Manual, p. 34.

* $p < .01$.

** $p < .001$.

relationships. They have been described as valuing social and religious achievement.

The profile of the pastors' group with the Artistic scale second indicated that pastors have artistic, musical, and literary interests and also value having a philosophy of life. The rank order of the two remaining scores obtained by the pastor sample support the description of the Social personality type for this group.

It was not surprising that pastors scored lowest on the Realistic and Conventional scales. Unlike pastors, the realistic type person tends to be mechanically oriented with low social interests and has an aversion for problems requiring a sensitivity for feelings. The Conventional type person achieves his goals through subordinate roles and by conforming and ordering his life according to prescribed ways of behavior. The scores obtained by the clergymen sample on remaining five VPI scales (Self-Control, Masculinity, Status, Infrequency, Acquiescence) further substantiated the SAIE pattern.

Summary. Pastors may be described as most sensitive to personal, humanitarian, social, and emotional influences. They are least sensitive to materialistic, influences and roles which require structured, conforming behavior. Such a description is, of course, a generalization, but it is interesting to note its similarity both to persons in the pastoral ministry and to the descriptions of desired characteristics of persons entering the ministry (Department of Ministry, 1969).

Most of the work of the average pastor is in direct relationship to persons and their needs (Social), but it also includes creative and innovative leadership (Artistic), study and teaching (Intellectual), and administration (Enterprising). Hence, the picture of clergymen (SAIE) generated from Holland's (1966) theory and the VPI appears to have some merit.

REFERENCES

- Department of Ministry. *The christian ministry/A challenge*. New York: Department of Publication Services, National Council of the Churches of Christ in the USA, 1969.
- Holland, J. L. *Manual for the Vocational Preference Inventory*. Consulting Psychologists Press, Palo Alto, California, 1965.
- Holland, J. L. *The psychology of vocational choice*. Waltham, Massachusetts: Blaisdell Publishing Co., 1966.
- Holland, J. L., Whitney, D. R., Cole, N. S., and Richards, J. M.,

Jr., *An occupational classification for research and practice*. ACT Research Report No. 29, Iowa City, The American College Testing Program, 1969.

Schuldt, D. L. Men leaving the pastorate: Social and psychological factors involved in career change of United Methodist Ministers. Unpublished Masters thesis, The University of Iowa, 1970.

BOOK REVIEWS

MAX D. ENGELHART, Editor
Duke University

HENRY MOUGHAMIAN, Assistant Editor
City Colleges of Chicago

- Aiken's Psychological and Educational Testing.* ROBERT M. COLVER 1031
- Bloom, Hastings, and Madaus' Handbook on Formative and Summative Evaluation of Student Learning.* WARREN G. FINDLEY 1033
- Cooley and Lohnes' Multivariate Data Analysis.* ROBERT M. PRUZEK 1036
- Edwards' Probability and Statistics.* JAMES A. WALSH 1039
- Thorndike's Educational Measurement.* NICHOLAS J. ANASTASIOU 1040
- Turney and Robb's Research in Education: An Introduction.* GERALD M. GILLMORE 1044
- Wittrock and Wiley's The Evaluation of Instruction.* JAMES R. SANDERS 1047

Lewis R. Aiken, Jr. *Psychological and Educational Testing*. Boston; Allyn and Bacon, 1971. Pp. vi + 346. \$9.50.

This book is designed to serve as a source book of information and procedures for counselors, teachers, and other persons concerned with testing and as a textbook for students in these areas. This reviewer feels that this book does adequately meet some of these objectives, but is inadequate in meeting others.

The book is basically organized in three general sections. The first section, Chapters 1 to 3, dealing primarily with background and methodology in testing; the second section consisting of Chapters 4 to 7 are primarily concerned with measurement in the cognitive field and the third section consisting of Chapters 8 and 9 emphasizing affective measurement.

Chapter 1, of the book, devotes itself to background and resource information for testing. The first section of the chapter gives a very brief, but sufficiently adequate overview of the history of measurement. It is felt by this reviewer that the section on "Sources of information" is too brief. The best that a textbook in this general field can hope to achieve is an overview of standardized tests that are available with encouragement to the readers to seek more detailed information in specific references devoted to standardized tests. A table of contents of the *Sixth Mental Measurements Yearbook*, quoted on page 7, does give the student knowledge of the various fields for which standardized tests are available.

According to the preface "no previous exposure to statistics is assumed." Yet, it appears that this book in its section on statistical methods in testing tries to go too far too fast. For example, by the sixth page of the discussion on statistical methodology the summation operation is described including the use of sigma complete with superscripts and subscripts. Four pages later the reader is asked to calculate a coefficient correlation based on summation operations and is given a formula with seven summation signs. A passing glance at that formula would probably discourage most of the readers for whom this book is designed and appropriate.

Chapter 2 concerns itself with the preparing, administering, scoring, and evaluating of tests and test items. This chapter is a very thoughtful and concise discussion of the problems and techniques in test preparation. It appears to be an excellent background for understanding the problems of test preparation, but does not appear

to be adequate for instructing the classroom teacher in the preparation of classroom tests.

Chapter 3 concerns itself primarily with the matter of the characteristics of satisfactory measuring instruments such as the questions of reliability, validity, standardization and norming. This material appears to be very adequately covered and at a level of sophistication appropriate for the group for which this book is designed. One particular aspect that was appealing to this reviewer was a brief, but very adequate, discussion of expectancy tables and a nomograph for predicting grade point averages. This type of applied material is especially appropriate for the audience to whom this book is directed.

The next two chapters are primarily an overview and directory of standardized achievement tests and individual and group tests of intelligence. Here the author has made what appears to be an excellent and very up-to-date selection of the more widely used tests in these areas with a brief annotation and evaluation of these tests. The material presented here gives a satisfactory introduction to the use and purposes of tests of this nature and the annotation is appropriate for the non-professional test user who wants to find some information about the more common tests used in the schools.

Chapters 7, 8 and 9 are similar to the two chapters mentioned above except that they cover the tests of special abilities, measures of interest, attitudes, and personality. As in the two previously mentioned chapters these three chapters serve as an excellent basic quick reference to up-to-date evaluation in these areas.

Research and theories on general intelligence is the subject matter of Chapter 6. Again, it appears that this is a superior overview and review of this area and at a level of sophistication appropriate for the individuals for whom the book is designed. It does appear that sufficient distinction between the ratio IQ and the deviation IQ is missing throughout all discussion of intelligence and intelligence measurement in the book. The final chapter on current issues and developments is an excellent summary of the modern day problems in testing and a very realistic forward look to what could or might be happening in the area of testing.

Overall the greatest strength in this book lies in well written up-to-date annotation and summary of the more common types of assessment devices currently in use in the schools. This does provide a valuable quick reference for the nonprofessional test user in assisting such persons in understanding the results obtained from testing programs as well as assisting in their communication with the professional test users.

A useful supplement to the text is a *Study Guide* containing chapter summaries, self-testing, exercises, lists of terms, and names of tests. An *Instructor's Manual* composed of parallel readings, lists

of audio-visual materials, and test items for each chapter is also available.

The reviewer can recommend this book as a textbook or a source book for background information or training of the non-professional test user such as the classroom teacher and the school administrator who need to have understandings in this area.

ROBERT M. COLVER
Duke University

Benjamin S. Bloom, J. Thomas Hastings and George F. Madaus.
Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill, 1971. Pp. iii + 923. \$11.95.

The scope and depth of treatment given to educational evaluation in this volume accounts for the delay in its publication. Quite wisely and courageously the authors relate evaluation to developments in curriculum and instruction and to the whole changing social setting in which evaluation is to be applied.

The organization of the book is reminiscent of the 1945 NSSE Yearbook, on *The Measurement of Understanding*. Substantial introductory sections of general application (280 pages) are followed by a doubly long section of eleven separately authored chapters devoted to statements of objectives and their evaluation in pre-school education, language arts, secondary school social studies, art, science, secondary school mathematics, literature, writing, second language learning, and industrial education. The area of preschool education is treated in two chapters, one devoted to evaluation of socio-emotional, perceptual-motor, and cognitive development, and the other covering early language development.

Although the authors properly point out that this is a handbook, hence not to be read cover to cover, the reader may well ponder the opening chapter giving their "View of Education." They forthrightly declare a belief in the fundamental teachability of all children, arguing the considerable modifiability of social-class-linked learning factors of standard language development, motivation to secure maximum education, willingness to work for teacher approval and/or long-term goals, and acceptance of school learning tasks "with a minimum of rebellion." With evaluation turned from selection to development, a modern version of the "plan-test-teach-test-plan" model gives a place for not only fuller specification of instructional objectives in terms of behavioral outcomes, but insertion of formative evaluation and feedback as intervening steps in the "test-teach-test" sequence. Their argument for the primacy of structure of the learning process over structure of a subject seems moot in that there may be more need among their readers for those

concerned with process to be attentive to subject structure than vice versa in the present cycle on that issue.

Learning for mastery is given the full treatment it deserves. This reviewer counts himself among those strongly influenced and helped by this approach in his instruction. Life is a work-limit rather than a time-limit situation largely, and giving students as long as they need and are willing to spend to master the content of most courses is productive in student learning. This is true regardless of whether the learning units have been or can be broken down all the way into blocks or elements for individual mastery. Cooperative effort by teacher and student to achieve fixed goals is an integral part of mastery learning, fitting it conceptually into the "View of Education" already presented.

The second major section on "Using Evaluation for Instructional Decisions" might be subtitled "The Humane Use of Tests." The need for summative, certifying or grading evaluation is not blinked, but it is put in a framework of teacher as helper to mastery by subsequent chapters on "Evaluation for Placement and Diagnosis" and "Formative Evaluation." These latter chapters expand to full treatment their introduction earlier in the discussion of formulating objectives. One wonders whether the present shift from summative and predictive evaluation may not some day make it more natural to present these topics in their proper chronological sequence of placement, diagnosis, formative feedback, and summative evaluation.

A third major section (105 pages) relates evaluation procedures to the several categories of the *Taxonomy of Educational Objectives*, Handbook 1—Cognitive Domain and Handbook 2—Affective Domain, by Bloom, Krathwohl, et al. It begins on a note of wholesome respect for knowledge outcomes, noting how these outcomes still are basic in the best curriculum outlines and guides. The logic of the several levels of cognitive outcomes: knowledge, comprehension, application, analysis, synthesis and evaluation is clearly spelled out and well illustrated with items from college and academic high school levels. Readers will generally be concerned with use of this framework in their subject areas, so may be directed to the separate chapters of Part 2 for illustrations more directly useful in preparing evaluation devices for their courses. One could wish for a broader set of references at many points. Creativity is discussed under "Synthesis" without mention or listing references to Guilford or Torrance. A warning against narrow interpretation of "Evaluation" as including mere preference is given without reference to the definitive research of Kripp and Stoker. And one misses any reference to the problem of presenting test exercises to young, slow, or foreign students less adept in the nuances of language taken for granted in the illustrations.

Chapter 10 on "Evaluation Techniques for Affective Objectives" presents a strong case for including evaluation of the achievement of such objectives. In response to the oft-voiced fear of brainwashing or invasion of privacy through grading of affective behavior of children, the authors point out that no summative grades need be given (or recorded) individually. Rather, formative evaluation may be used to feed back constructive guidance to individuals, while anonymous group data may be used formatively and summatively to evaluate the curriculum and instruction. The very real problem of enlisting individuals in their own improvement and the problem of the "socially desirable" answer remain, but the "intangibles" need to become thus much more tangible, rather than left as vaguely hopeful long-term outcomes.

Short, efficient chapters are devoted to developing the technology of evaluation systems and emerging developments in evaluation. In the first of these, total evaluation systems are conceived as basic to helping the teacher help students learn by organizing a supporting technology and specialists to give leadership and guidance in its use. The second chapter seems unduly eclectic, quoting extensively from National Assessment, but omitting the systematic work of Guba and Stufflebeam and the whole concept of accountability.

The subject chapters deserve separate review by separately competent specialists. Suffice it to say here, that this reviewer found the two chapters on preschool education especially rich with detailed illustration and definite comment. Kamii's chapter has the special merit of being based on an experimental program in which a Piagetian curriculum is struggling to be born. Cazden's chapter happily concentrates on new insights of psycholinguistics as applied to the fundamental problem of compensatory education; her comments on the use of specific standardized tests are particularly cogent.

It is also significant that "Evaluation of Learning in a Second Language" is presented by one who has taught English as a second language and sees foreign language instruction and its evaluation in that context. Some of these insights, combined with the observations on preschool language development, might illumine the more traditional breakdown into language arts, literature and writing in the typical English curriculum.

Unfortunate omissions are chapters presumably originally planned in elementary social studies and mathematics. The chapter on secondary school social studies suffers most because of the still prevalent tendency to equate social studies at that level with history and government. Elementary social studies is a ferment of disciplines, including rediscovery of geography. The chapter on secondary school mathematics will interest many because of the adaptation of the Bloom taxonomy to fit mathematics objectives. The original taxonomy has had the virtue of showing that one reason mathema-

tics has been so hard to many students is the failure until recently to reward anything below problem-solving as achievement worthy of noting.

At one point in writing this review, your reviewer was prepared to organize the AAHSB (Association against Heavy Slippery Books). Not only is such a book difficult to read in the bathtub, but its weight discourages its use as a textbook. However, as he contemplated the cost of printing a main volume with paperback chapters on the specific subjects, it became obvious that a single-volume edition was essential if costs were to be kept in bounds as they have been. So, here's to a truly elegant volume, in the best sense of that adjective. It is a reference that should stand long despite the shifting sands of curriculum and society.

WARREN G. FINDLEY

University of Georgia

William W. Cooley and Paul R. Lohnes, *Multivariate Data Analysis*, New York: John Wiley & Sons, 1971. Pp. x + 364. \$9.95.

Multivariate Data Analysis is an appealing title for a book. Especially for those persons who are familiar with John W. Tukey's essays on data analysis, this title is apt to conjur up visions of a revolutionary text on multivariate methods. If contents of a book were to adhere closely to Professor Tukey's thinking they would, in relation to contents of other extant books, be more strongly and more persistently oriented towards the scientist's working questions—such as: Given my data, just how have I added to my knowledge of the phenomena under study? How might my data be more productively gathered and analyzed at successive stages of research? How might my motivating questions be most effectively sharpened for the collection and analysis of further data? Etc. Tukey has argued for several years that there is a place for a science called data analysis, where subject-matter questions are held to be preeminent, where formal methods and models for inference are subordinated to informal ones, where mathematical models are used for guidance in the conduct of research but where these models are never believed in, except as aids in the analysis of questions which bear on the phenomena or data themselves. This approach to quantitative thinking has been hotly debated among methodologists in numerous disciplines and, in my judgment at least, deserves to be aired in systematic textbooks.

Irrespective of the title, and numerous specific Cooley and Lohnes' endorsements of Tukey's thinking, this is *not* a revolutionary book on data analysis. Both the topics and their treatments are essen-

tially conventional. The text consists of three Parts which have been further divided into 13 chapters. Part I (two chapters) includes a general overview and a very brief introduction to vectors and matrices. Part II (five chapters) is entitled "Studies of a Single Population," with chapters covering partial, multiple and canonical correlation as well as principal component analysis. Part III (six chapters) is entitled "Multiple Population Studies"; multivariate analysis of variance, discriminant function analysis (and classification procedures), and general strategy considerations are offered. The entire approach is intended as introductory; the reader is assumed to be an applied researcher who is at least familiar with classical univariate procedures. There are 60 pages of FORTRAN program listings as well as a brief FORTRAN primer "to whet your appetite and give you courage" (p. 26). For persons who have seen the authors' 1962 volume, *Multivariate Procedures for the Behavioral Sciences*, the present book will appear essentially as a revised and expanded version.

On the positive side, the authors have clearly improved on their 1962 effort. This version is physically attractive in both format and style of print and suffers from few misprints (although some minor slips were made in bold print expressions on pp. 59, 141 and 177 and a table is misaligned on p. 134). Several of the expository and matrix-based discussions of methods are reasonable, as far as they go, and the associated FORTRAN programs rely on many of the newer developments in numerical analysis. Taken together, the chapters of this book provide several examples and substantial commentary as to how multivariate methods may be used in behavioral and social science research; perhaps for many people this is enough to merit a recommendation.

Unfortunately, a number of points must be made on the negative side. Some of the problems seem to require little commentary. Examine the following statements in relation to one another:

"Most of our procedures are concerned with estimating or making inferences about parameters of a m [multivariate] n [ormal] d [istribution]" (p. 1).

". . . for the most part the examples are from surveys. The emphasis on heuristic rather than hypothesis testing uses of the procedures follows from this preoccupation with survey sciencing" (p. 7).

"By a model we mean the basic matrix algebra specification of a procedure for analyzing data" (p. 10).

"This book is not concerned very much with distribution theory, nevertheless it is worthwhile to take a look at some of the properties of a multivariate normal distribution . . ." (p. 35).

Despite the disclaimers, dozens of t -, F - and Λ -statistics are to be

found, and some of the chapters contain several pages on tests of significance. Strange.

Other anomalies:

"We do not undertake . . . to fit higher order polynomials because it seems to us that even a finding of significance for a cubic term would so discourage the behavioral scientist that he would want to change his method of scaling one of his constructs, or change his research design, or his line of work" (p. 79).

"A living human being is a highly integrated system, all the overt characteristics and behaviors of which are interrelated. *If we want an uncorrelated vector variable, we have to construct it by transforming the data*" (authors' italics) (p. 97).

Further problems are to be found in the chapters on component analysis. At several points the authors imply that statistical inferential tests are available for testing hypotheses on the basis of characteristic roots and vectors of correlation matrices. While it is true that certain tests are available for (Thurstonian) common factor methods (using maximum likelihood statistics) it has not been possible to develop formal hypothesis testing methods for components methods. Other confusions in these chapters have to do with improper use of the term "communality" (see especially, p. 150); strictly speaking, this term has no relevance to component analysis. Also no distinction is made in the discussion of "rotation" methods, between primary and reference axis systems. This results in confusions between pattern and structure matrices.

The book can be faulted for a total lack of sustained methodological themes. A major example has already been noted—that the authors see themselves teaching Tukeyian data analysis, but that they continually fall back on classical significance tests and associated baggage; and there is no mention of graphical plotting methods or general procedures for improving the *fit* of models to data. Tukey calls "fitting" the "workhorse of data analysis." Another problem in this context is that while all the methods are special cases of the general (multivariate) linear model, it is almost impossible from this presentation to see most of the basic interrelationships. Besides missing several opportunities to synthesize methods, it is rare to find discussions showing how a single method may have different uses, depending on the investigator's questions. It does not help to find nearly all examples from the survey files of PROJECT TALENT.

While this book may meet certain limited purposes, it should be clear that any recommendation must be substantially qualified. The title of the book seems particularly misleading. Forthcoming books by Tatsuoaka (*Multivariate Analysis for Educational Research*; John Wiley and Sons), Bock (*Multivariate Statistical Methods in Behavioral Research*; McGraw-Hill) and a recently

published book by Rummel (*Applied Factor Analysis*; Northwestern University Press) may be worthy alternatives to this one.

ROBERT M. PRUZEK

*State University of New York
at Albany*

Allen L. Edwards. *Probability and Statistics*. New York: Holt, Rinehart and Winston, 1971. Pp. xvi + 257. \$8.50.

This moderate-sized text for introductory courses in applied probability and statistics is meant for students with adequate grounding in algebra but none in calculus. Its basic emphasis "is on proving equations and theorems that the student is ordinarily asked to take on faith . . ." Proofs are almost exclusively in terms of discrete variables in order to avoid the necessity for calculus. In addition to use as a text, Edwards recommends the book for supplementary reading in applied courses. It is obvious that this is the real purpose for which the book was meant, since the mode of presentation is relatively formal, the amount of exposition and linking material is generally minimal, and attention to direct applications is practically nil.

In practice, this book could be used to supplement an intuitively-oriented text or lectures emphasizing intuitive ideas and applications, by providing formal proofs of assertions. For such use it has several advantages. Foremost of these is the clarity and consistency which characterize Edwards' writing style. Assumptions are always clearly and explicitly stated, proofs are developed in an orderly step-by-step fashion without mysterious jumps, and terminological and notational conventions are defined and then adhered to throughout.

Of comparable importance are the range of topics and the care given to their selection. The book begins with the algebra of samples and proceeds to the ideas of sample spaces and probabilities defined upon them. Counting procedures follow and then discrete random variables and their expected values. After this Edwards develops the properties of random samples drawn from finite populations without replacement; binomial variables and ranked variables are then treated, followed by the Poisson, normal, chi-square, Student's t , and F distributions. The three final chapters deal with expected mean squares in one-way anova, power, and confidence intervals, respectively. Almost all topics which might be contained in the usual introductory course are covered. In addition, the chapters following the one on sampling without replacement from finite populations are relatively self-contained and could be used more-or-less independently.

Three aspects of this coverage seem especially good to me. First, the distinction between population and sample standard deviations is made and maintained without fuss and without confusion. Second, the treatment of sums of simple random variables is thoroughly done, making possible an easy transition to notions of linear combinations in later courses on measurement or factor analysis. Third, the Poisson distribution is developed and linked to other distributions in an understandable way.

Areas of strength notwithstanding, the book has several weaknesses. It would be more in line with contemporary usage to describe experiments in the behavioral sciences as "modeled by" or "represented as" random experiments rather than saying they "are random experiments" (p. 26). Such a statement seems likely to blur what is, for introductory students, the already fuzzy line between scientific method and mathematical models.

A second fault, and one shared by most writers of statistical texts, is the failure to attribute important features such as the handling of particular topics to the men who originated them. Edwards' treatment of probability owes a great deal to Feller; most chapters on probability in most texts on methods owe similar debts to Feller and the time has come to acknowledge them.

On the whole, this text is well done. However, the purposes for which it was designed should be kept in mind by a prospective user. By itself it contains too little expository material to serve as a primary text for an introductory applied statistics course. To be used as such, a lecturer would have to provide both the intuitive background for the material and also illustrative applications. Used simply as a source of supplementary material on the mathematical underpinnings of applied statistics, it should prove most satisfactory.

JAMES A. WALSH

Iowa State University

Robert L. Thorndike. *Educational Measurement*. (2nd. ed.) Washington, D.C.: American Council on Education. Pp. 768. \$15.00.

The first questions to be asked of a volume of this nature is who is its intended audience and what is its general usefulness. This volume falls somewhere between an encyclopedia and a mechanics manual, it contains compendiums of current facts, step by step how-to-do-its, and theoretical discussions of varying levels of sophistication which range from very exciting to disappointing. The problem of single edited multi-authored, omniscient volume is its unevenness. As encyclopedias vary in their descriptive excellence so do reference books of this nature.

The reviewer faced the same problem as the editor—how to evaluate such a broad coverage of the field. The list of authors and

those who read or contributed to the various chapters is a Who's Who in test and measurement. This reviewer, in need of consensual validation, turned to his colleagues, Clint Chase, Gerald Bracey, Richard Pugh and Sydney Mifflin, who read sections of the book and discussed them with him. The reviewer, however, takes full responsibility for what appears here.

The appropriate audience for this book is a diverse one and unhappily, to this reviewer one which does not include teachers and administrators. There are excellent chapters for the theoretician and the advanced student and some excellent sections for the test item writer, the printer, and the clerk in a testing department. However, a teacher turning to this reference work would find that the chapters which might be useful are overly drawn out, full of platitudes and display a surprising lack of sophistication about children and schooling. For the most part the articles contain psychometric theory by educational psychologists who are apparently much less versed in knowledge of schools, schooling and children.

The first chapter of the book begins with an excellent overview by Robert Thorndike of the changes that have taken place during the past twenty years in the field of test and measurement. He places emphasis on the role of the computer and data processing in test development. Thorndike's emphasis in this chapter is one of adequate data collection. He holds that test producers can no longer justify shortcuts now that the rapid analyses features of computers are available to the test constructor. To this reviewer's taste, Thorndike's discussion of the political and social problems of testing are only superficially summarized. The first section of this book covers *Test Design, Construction, Administration and Processing*. Chapter Two, *Defining and Assessing Educational Objectives*, is a disappointing chapter. The authors focus clearly on the need of test constructors to define pupil behavior and necessity of attempting to maximize the probability that the test be a measure of student learning. However, limited space was given to discussion of the critical issue of values and how the test developer selects objectives, i.e., how does the developer attempt to meet societal or humanistic needs. Lindquist's brilliant chapter in the 1951 *Educational Measurement (Preliminary Discussion in Objective Test Construction)* is quoted but his ideas are not developed nor applied as fully as one would hope, given the concerns of the 1970's.

Chapter Three, *Planning the Objective Test*, is a compendium of folk lore and unvalidated common sense assumptions. This is a very long chapter which could be greatly shortened by more concise writing. High points of the chapter are the editor's excellent note on the problem of guessing and the author's discussion on practical issues of reliability.

Chapter Four, *Writing the Test Item*, should be useful to the clerk in a test production department. It contains many examples of items and the general suggestions for writing items are excellent. However, helpful though the chapter might be, it is unfortunately a composite of a multitude of unnecessary platitudes, i.e., "a good test is composed of well written items (81)."

A more important criticism of this chapter is the author's suggestion to include test items to force teachers to attend to areas of the curriculum that they are ignoring even though the psychometric properties of the items may not be fully acceptable. This suggestion seems to this reviewer a very questionable procedure, both psychometrically and ethically.

Chapters 5-8 concern *Gathering, Analyzing and Using Data on Test Items, Reproducing the Test, Test Administration and Automation of Test Scoring, Reporting and Analyses*. They will be undoubtedly useful to the test department of a commercial firm or large school district.

Chapter 6, *Reproducing the Test*, should be particularly useful to the commercial printer or those who give instructions to the printer.

Part two of the book covers *Special Types of Tests*. Chapters include *Performance and Product Evaluation, Essay Examinations, Prediction, Educational Outcomes*. All of the authors appear to be conversant with measurement techniques but weak on psychology. Some of the chapters in this section get carried away with rigidly applied procedures and rules as if our measuring instruments were perfect. Perhaps if all the authors in this section had read Cronbach, Jones and Davis, included elsewhere in the volume more levity could be applied to the current state of the art involving design and construction.

The strongest section is section three, *Measurement Theory*, particularly Lyle Jones's *The Nature of Measurement*, Stanley's *Reliability*, and Cronbach's *Test Validation*. However in the same section, Angoff's otherwise excellent chapter on *Scales, Norms, and Equivalent Scores* is exhaustive beyond the point of relevancy and Cooley's *Techniques for Considering Multiple Measurements* is hardly above an annotated bibliography. Stanley's chapter is well written, makes complex ideas seem simple, deals exhaustively with reliability theory, and avoids cookbook procedures. However, readers of this chapter will have to have a background in statistical theory in order to apply its procedures. This reviewer would have profited from Stanley's including more discussion on "weak" and "strong" true score theory, however this is a minor point.

Cronbach likewise writes clearly about complex issues. His presentation of construct validity is outstanding. He discusses his departure and agreement with Loewinger's earlier stand. He speaks

to the ultraoperationalist and non-behavioralist as well. He and Stanley have the ability to write simple statements that assist the reader center on the basic issues of test theory, i.e., "one does not validate a test but the interpretation of data arising from a specified procedure (447)." Cronbach is immensely quotable.

Angoff's chapter is very comprehensive in scope and could be used in production endeavors. He includes a section on sampling and design for methods of equating test forms which should prove to be exceedingly useful to the test producer. Angoff's examples are frequently made to physical measurements and the analogies are somewhat debatable. He makes the analogy between a scale of weights and a scale of typing ability. Although Angoff's analogy is helpful in parts, applying the analogy to a multi-factored human skill is a debatable practice.

Jones's chapter on *The Nature of Measurement* covers a breadth of topics usually covered in test theory courses. He is readable. Jones stresses appropriately that the necessary prerequisite before measurement can be made is to define the attribute in quantifiable terms that contain meaning. This is an excellent point and he holds to it throughout his chapter. However, when he gives a case study to clarify the conception and perception of attributes he uses the example of the case history of length rather than an example from an educational setting. Likewise his example in his excellent section on classification and rankings are drawn from other than education and psychology. That is this reviewer's major critique of this chapter is of the tendency to use non-educational examples to make a point or to compare physical measurement scales for the purpose of developing a frame of references for the state of the art of educational measurement. Jones uses non-educational examples frequently in his otherwise excellent chapter.

For example, Jones's use of examples outside of education is entertaining in the case history of length, irrelevant in the discussion of classification and rankings and distracting in the section of unit of measurement. Jones's final section contains a discussion of Campbell's classic theses of the 1920's and the alternatives behavioral scientists have produced. Jones, in this section, returns to his major theme: it is the meaningfulness of the empirical counterpart which are important not what you call it.

Section Four, *Application of Tests to Educational Problems*, is a strong section of what has yet to be a well developed field, i.e., evaluation. Glaser's and Nitko's discussion of the four activities of instructional design, i.e., analyses of subject matter, diagnoses of characteristics of the learner, design of instructional environment and evaluation of learning outcomes is a good one and well developed. Their discussion of instructional models are current and their discussion of norm versus criterion reference tests should be

useful to program evaluators. Davis's use of *Measurement in Student Planning and Guidance* is useful but should be read in conjunction with Hill's chapter, *Use of Measurement in Selection and Prediction* which is largely aimed at college selecting.

The Astin and Panos chapter, *The Evaluation of Educational Programs*, clearly defines evaluation as the collecting of information upon which to base a decision. However, in this reviewer's opinion they superficially overview the field. In addition, when they do draw upon the developmental needs of children they draw upon Bruner's 1961 work, Piaget's 1950 work and Erickson's 1950, even though Bruner's circa 1961 position was inconsistent with Piaget's circa 1950.

This book has several chapters that will serve as a useful reference both for advanced students in educational psychology, particularly those in advanced measurement courses. Rarely does the volume contain new information for the professional. For the beginning test developer it will provide some useful hints and suggestions. Its strength is its weakness; it covers so much that users will have to be very selective in what they recommend to their students. The strong points have been mentioned above; its weakness is that there are all too few educational psychologists who specialize in measurement that can bring to bear a knowledge of children and curriculum. Isn't it about time we in education demanded that the level of scholarship about measurement be matched with an equally high level of scholarship about children and schools?

NICHOLAS J. ANASTASIOU
Institute for Child Study
Indiana University

Billy Turney and George Robb. *Research in Education: An Introduction*. Hinsdale, Ill.: The Dryden Press, 1971. pp. xi + 320. \$6.95 (paperback).

This book is essentially a "how to do research" text. It was written primarily for use in the initial course in research methodology by students of education. The authors suggested that the book could also "... serve as a useful reference for classroom teachers, counselors, or administrators who are interested in doing research but need a 'refresher course'" (p. vii). Twelve chapters cover the research process from selection of a researchable problem to writing the final report.

In the introductory chapter, an attempt is made to define and explain research and the scientific method. Inductive and deductive logic are discussed briefly. Research is categorized into basic and applied, as well as into historical, descriptive, and experimental.

As an introduction this chapter seems somewhat brief and fragmented. However, most of its topics are dealt with at more length subsequently.

"Selection and Evaluation of a Problem" is the topic of chapter 2. The authors display good common sense in this section, and do well in treating this important aspect of research which does indeed often plague fledgeling researchers.

In discussing "The Research Proposal" in chapter 3, Turney and Robb included the elements that one expects to find in a well-written proposal. However, the order of a proposal is presented as being quite rigid and nonadaptable. One should be allowed more freedom to tailor the form of a proposal to fit a specific problem than the authors seem willing to permit. Throughout this chapter, relevant examples are used very effectively to illustrate points being made.

The next two chapters concern the use of the library. Chapter 4 is an adequate if tedious presentation of what is available in a library, while chapter 5 handles library use. Again, suggestions sound more like prescriptions, such as the specification of the exact format one should use for recording reference information on 3×5 cards. More emphasis could have been beneficially placed on the writing of nonpedestrian literature reviews.

In chapter 6 the three types of research introduced in chapter 1 are elucidated. The brevity with which the authors chose to write weakens this chapter. Descriptive research is limited primarily to surveys and case studies, with prediction studies, for example, never even mentioned. Experimental research is given a "once over lightly" focusing on field studies, field experiments, and independent and dependent variables. Designs containing other than one control group and one experimental group are neglected completely.

The next chapter is entitled "Analysis and Treatment of Data." Approximately 25 statistical topics are presented from frequency distributions and percentiles, through confidence intervals and *t* tests, to regression lines and Spearman's rho. In condensing most of the topics found in a complete statistical textbook into 32 pages, frequent misinterpretations, omissions, symbols and concepts used without definitions, etc., resulted. More importantly, the possibility of converting a naive student into a reasonably sophisticated user of those 25 techniques in that brief a space is highly questionable. Given the restraints, emphasis would seem to have been better placed on the understanding of data, data analysis, and the use of statistics. A list of references could well have been included for specific techniques.

The dual themes of chapter 8, "Factors Affecting Research Results," are what Cambell and Stanley (1963) have labeled internal

and external validity. Both subjects are handled interestingly and adequately. In chapter 9, a long and nicely representative list of paper-and-pencil data gathering devices is offered. Included is a description of the use and interpretation, as well as the limitations, of each. The sections on limitations do tend to be somewhat pessimistic in that techniques which can overcome specific limitations go unmentioned. The most surprising omission in this chapter is the absence of references to more extensive presentations of the various techniques.

The subject of chapter 10 is "Computational Aids for the Researcher." It is divided into two pages on desk calculators and nine pages on digital computers. Frankly, the section on computers is incredibly poor. About 90 per cent of this section concerns content which educational researchers do not need to know, and most probably do not really care about; namely, computer hardware. Furthermore, even the explanation of computer hardware is muddled. For example, core storage is described in terms of ferromagnetic rings and electrical charges of changing directions. However, the fact that all storage and processing is done in binary is kept secret. Bytes and words are mentioned, but bits are not.

Most universities which support graduate research have a computer installation with user oriented "canned" programs which meet the needs of most researchers, especially beginners. Thus, this chapter on computers would seem more sensible if it contained some orientation to the workings of a computer facility as they relate to a user, and to the use of existing programs. This is the aspect of computers which is of vital importance to the book's intended audience. In spite of this, Turney and Robb devoted only two sentences to program libraries and one to the workings of a computer facility.

Chapter 11 contains specific rules on the nitty-gritty of "Reporting Educational Research," such as the proper use of pronouns, footnotes, and references. Following the text of this chapter are 28 pages of examples which illustrate the preceding principles. Two aspects of this chapter are curious. First, no direct reference is made in the text to the appropriate examples which follow. Thus, when a student is reading about footnoting, for example, he is not referred to the subsequent examples but must find them himself. Secondly, the standards for publication of the American Psychological Association (APA) are neither used nor even mentioned, and are often violated. Given the number of educational journals using the APA format, this is not a trivial oversight.

Chapter 12 is the book's greatest strength. A research proposal and a research report are presented, and each is expertly critiqued by Linda Mitchell Crocker of the University of Florida. The pedagogical value of this section is moderated because each ex-

ample is presented in toto, followed by its critique. Were the specific comments of the critiquer to appear contiguous with the aspect being commented upon, possibly in a double-column presentation, the result would be much more effective. Also, the research proposal is criticized for not having a title when, in fact, it has; thus, indicating some editorial carelessness.

Overall, *Research in Education: An Introduction* by Turney and Robb can be described as brief, not very sophisticated, and superficial. Some parts are poorly planned and written. Therefore, use of this book as an exclusive text for a course is questionable. However, it could serve as a supplement, possibly to provide a student with a fairly quick overview of the research process. To commend it, the book is generally readable and contains many examples. The closing critiques are especially worthwhile. In sum, it is probably fair to say that the authors attempted to present too many topics in too short a space. In trying to serve two masters, brevity and comprehensiveness, they tended to lose sight of their ultimate masters, students.

REFERENCE

Campbell, D. T. and Stanley, J. C. Experimental and Quasi-Experimental Designs for Research on Teaching. In Gage, N. L. *Handbook for Research on Teaching*. Chicago: Rand McNally, 1963.

GERALD M. GILLMORE

University of Illinois

(Champaign-Urbana Campus)

M. C. Wittrock and David E. Wiley (Eds.) *The Evaluation of Instruction*. New York: Holt, Rinehart and Winston, 1970. Pp. xiii + 494. \$4.95.

Recent evaluation mandates by federal legislation, criticisms from the public sector which suggest that educators have not accounted for the resources they have used up, teacher demands for a greater voice in school planning and administration, and the appearance of a large number of instructional alternatives generated by curriculum projects, commercial developers and management consultants have all created an unmet need for methods for planning and conducting educational evaluation studies. With evaluation being the hot topic that it is today, a text with the title, *The Evaluation of Instruction* must arouse great expectations on the part of the educational practitioner. Evaluation specialists do not have a set of guidelines parallel to those provided to the research specialist by works such as the Campbell-Stanley treatment of experimental design. It is unfortunate that these expectations are not completely

met by the Wittrock and Wiley book, although some excellent logical frameworks are provided.

The text is primarily the product of a symposium sponsored by the UCLA Research and Development Center for the Study of Evaluation held at UCLA on December 13-15, 1967. The symposium papers and discussions are supplemented by four papers on causal models which were originally published elsewhere. The volume is divided into seven sections.

The first section, an introduction by Wittrock, reflects one theme of the text when Wittrock argues that in evaluating instruction one is usually trying to estimate cause-and-effect relations in non-experimental data to make judgments and decisions about instruction. This statement begs the question of whether causal inferences are central to evaluation studies. The purpose of any evaluation study is to determine the worth of some phenomenon. With that purpose in mind the evaluator will identify a set of questions about the phenomenon to be answered by the study. Certain information needs will follow directly from the questions. If the needed information is causal or comparative information, the use of true experimental designs or the correlational techniques suggested in the appendix of the Wittrock-Wiley volume would be appropriate. If, as is often the case in educational evaluation, no causal inferences are required to determine worth, then the evaluator should in no way feel obliged to estimate cause-and-effect relations. Many excellent evaluation studies have been conducted without any explanatory attempts.

The second part, entitled, "Theory of Evaluation of Instruction," includes papers by Benjamin Bloom and Robert Glaser, formal comments of the Bloom paper by Michael Scriven, Gene V Glass, and J. P. Guilford formal comments on the Glaser paper by Robert Stake and Arthur Lumsdaine, and transcripts of open discussions. The open discussions are recorded in all of the following sections as well. Bloom delineates what he calls three approaches to testing (measurement, evaluation, and assessment) and attempts to consolidate the three approaches into one theory of testing. The attempt to push evaluation under the heading of testing is eloquently refuted in the discussion of his paper. Glaser suggests a general model of instruction in his paper and lists issues for evaluation and measurement that are inherent in the model. It becomes apparent from Glaser's discussion that the methodology of measurement may need more work than that of evaluation.

The third section entitled "Instructional Variables," contains a paper by Robert Gagné and comments by Richard Anderson, Leo Postman, and John Bormuth. Gagné, in one of the more significant presentations of the symposium, argues skillfully that two primary criteria of measurement are distinctiveness (of measurement opera-

tions for any one inferred entity) and freedom from distortion (noise). The design of distinctive measurement must involve a two-stage operation. Anderson extended Gagné's major points to emphasize the importances of systematic analysis of test stimuli as compared to instructional stimuli. The Gagné and Anderson papers should be required reading for any student of psychological measurement.

The fourth section contains a paper by Dan Lortie, and comments by C. Wayne Gordon and N. L. Gage. The section is entitled "Contextual Variables." Lortie presents a concise, but convincing argument about the diversity of evaluation roles that are resulting from innovation and large-scale organizational change. The paper is essential reading for all evaluation students.

The fifth section, entitled "Criterion Variables," includes papers by Samuel Messick and Marvin Alkin. Comments on the Messick paper are provided by Paul Blommers and Leonard Cahen and comments on the Alkin paper are given by Marvin Hoffenberg and John Bormuth. Messick provides a detailed argument for a focus on cognitive styles, and a lesser argument for a focus on affective reactions, in evaluation studies. It is unfortunate that a more systematic analysis of the problem of specifying and measuring unintended outcomes of educational programs was not attempted, since the title of Messick's paper suggested a more comprehensive discussion. Alkin suggested a cost-effectiveness model as a tool for educational evaluators. The attempted model suggests a healthy movement toward investigating the utility of techniques from disciplines (e.g., economics) other than psychology and applied mathematics for use in educational evaluation.

The final section of the symposium, entitled "Methodological Issues," contains papers by David Wiley and Martin Trow. Comments on the Wiley paper are provided by Chester Harris and Theodore Husek and comments on the Trow paper are given by Eugene Litwak and David Nasatir. Wiley's paper, limited by a very narrow definition of evaluation, is most valuable for his discussion of specific analysis techniques available for use by evaluators and researchers. Trow discusses problems in evaluation design in higher education.

The last section of the book, an appendix, contains two papers written by Herman O. A. Wold, a paper by Otis D. Duncan, and a paper by A. H. Yee and N. L. Gage. The focus of the papers is toward teasing causal inferences out of nonexperimental data. This appendix was included, no doubt, to address the concerns contained in Wittrock's introductory section of the text.

With eminent scholars, such as those listed above, participating in the UCLA symposium, it would be highly improbable that no major contributions would be recorded in the Wittrock-Wiley vol-

ume. The comments by Stake, Glass, and Scriven and the papers by Lortie and Alkin should be on all reading lists for evaluation courses. The paper by Gagné and the comments by Anderson, Postman, and Bormuth should be required reading for students of measurement. The paper by Wiley and those papers attached in the appendix of the volume are important readings for students of research design and analysis, measurement and evaluation.

The volume fails to meet its promise in that many issues and few answers are provided for the practicing evaluator and that the promises of the book title and section headings are never fulfilled. These shortcomings are undoubtedly a function of the time at which the symposium was held. In 1967 much less had been written about the evaluation process than today. As Chester Harris remarked at the symposium, the area of the design and analysis of evaluation studies was actively changing and developing, and most of the participants would be hard pressed to predict the extent to which the issues would be resolved or reformulated in the near future.

It is interesting to look at the volume as a historical document, especially regarding the confusion of the symposium participants over what evaluation is and whether the evaluator actually has a role separate from the psychologist, measurement specialist, or researcher. If the symposium proceedings had been published in 1968 instead of 1970, the volume would have had considerably more influence on the definition of the emerging inquiry process. It should be noted, however, that, even today, few answers exist to the many issues identified at the symposium. It would be very desirable to continue to conduct periodically symposia of the type recorded in the Wittrock-Wiley volume with published proceedings available within six months after the end of each symposium. New members to the group, with fresh ideas, and representing other disciplines in addition to psychology and applied mathematics, could contribute greatly.

There is no doubt that the volume will serve well as a reference in evaluation, statistics and research design, measurement, and educational psychology as well as a valuable historical document to students of evaluation.

JAMES R. SANDERS

*Educational Research and Evaluation Laboratory
Indiana University*

INDEX FOR VOLUME 31

Abbott, Robert D. <i>A Factor Analysis of the CPI and EPI . .</i>	549
Abrahams, Norman (with Edward Alf). <i>A Significance Test for Biserial r</i>	637
Al Amir, Hudhail (with William B. Michael, Robert A. Jones, Calvin M. Pullias, Michel Jackson, and Valerie Goo). <i>Correlates of a Pass-Fail Decision for Admission to Candidacy in a Doctoral Program in Education</i>	965
Aleamoni, Lawrence M. <i>MERMAC Test and Questionnaire Analysis System</i>	777
Alf, Edward (with Norman Abrahams). <i>A Significance Test for Biserial r</i>	637
Allen, Dwight W. (with William P. Gorth and Aram Grayson). <i>Computer Programs for Test Objective and Item Banking</i>	245
Anderson, Edwin L. <i>The Use of the Common/Data Statement to Determine the Type of an Event in Simulation Studies</i>	771
Ayers, Jerry B. <i>Predicting Quality Point Averages in Master's Degree Programs in Education</i>	491
Baker, Frank. <i>Measures of Ego Identity: A Multitrait Multimethod Validation</i>	165
Bayroff, A. G. (with Carrie Wherry Waters). <i>A Comparison of Computer-Simulated Conventional and Branching Tests</i>	125
Bayuk, Robert J., Jr. (with Barton B. Proger, John R. McGowan, Lester Mann, Ruth L. Trevorow, and Edward Massa). <i>The Relative Predictive and Construct Validities of the Otis-Lennon Mental Ability Test, the Lorge-Thorn-dike Intelligence Test, and the Metropolitan Readiness Test in Grades Two and Four: A Series of Multivariate Analyses</i>	529
Berdie, Ralph F. <i>Self-Claimed and Tested Knowledge</i>	629
Bligh, Harold F. (with Joanne M. Lenke and Bernard H. Kane). <i>Cross-Validation of the Orleans-Hanna Algebra Prognosis Test and the Orleans-Hanna Geometry Prognosis Test</i>	521
Borich, Gary D. <i>Interactions among Group Regressions: Testing Homogeneity of Group Regressions and Plotting Regions of Significance</i>	251

Bower, Samuel M. (with Lester C. Shine II). <i>A One-Way Analysis of Variance for Single-Subject Designs</i>	105
Bowers, John. <i>A Note on Gaylord's "Estimating Test Reliability from the Item-Test Correlations"</i>	427
Bowman, J. Thomas (with Marvin E. Shaw and Frances M. Haemmerlie). <i>The Validity of Measures of Eye-Contact</i> ..	919
Brewer, James K. (with Edward P. Labinowich). <i>Computer Programs for Rank Analysis of Covariance</i>	295
Burg, Eldon (with John Follman and William Miller). <i>Statistical Analysis of Three Critical Thinking Tests</i>	519
Burnham, Paul S. (with Benjamin A. Hewitt). <i>Advanced Placement Scores: Their Predictive Validity</i>	939
Capra, J. R. (with R. S. Elster). <i>A Note on Generating Multivariate Data with Desired Means, Variances and Covariances</i>	749
Carter, Walter H., Jr. <i>The Probability of Misclassification of Students on Multiple Choice Examinations</i>	831
Cattell, Raymond B. (with Samuel E. Krug). <i>A Test of the Trait-View Theory of Distortion in Measurement of Personality by Questionnaire</i>	721
Centra, John A. <i>Validation by the Multigroup-Multiscale Matrix: An Adaptation of Campbell and Fiske's Convergent and Discriminant Validational Procedure</i>	675
Chissom, Brad S. (with Ralph Lightsey). <i>A Comparison of the D-48 Test and the Otis Quick Score for High School Dropouts</i>	525
Chissom, Brad S. (with Jerry R. Thomas). <i>Multivariate Validity of the Otis-Lennon Mental Ability Tests Primary I Level</i>	991
Clark, William H. (with Bruce L. Margolis). <i>A Revised Procedure for the Analysis of Biographical Information</i>	461
Clarke, Robert R. (with David A. Payne and Robert A. Wells). <i>Another Contribution to Estimating Success in Graduate School: A Search for Sex Differences and Comparison between Three Degree Types</i>	497
Costin, Frank. <i>Dogmatism and Conservatism: An Empirical Follow-up of Rokeach's Findings</i>	1007
Costin, Frank. <i>Hostility and Learning: A Follow-up Note</i> ..	1015
Cowan, Gloria (with S. S. Komorita). <i>The Effects of Forewarning and Pretesting on Attitude Change</i>	431
Cox, John A. (with Joseph P. Schnitzen). <i>Concurrent Validity of a Literature Test in Relation to Selection of Persons for Graduate Study in English</i>	485
Creager, John A. <i>A FORTRAN Program for the Analysis of Linear Composite Variance</i>	255

Crews, Sharon L. (with John R. Howell). <i>Eigenvalues and Vectors of Large Matrices on the IBM-1130</i>	263
Cureton, Edward E. <i>The Stability Coefficient</i>	45
Cureton, Edward E. <i>Reliability of Multiple-Choice Tests is the Proportion of Variance Which is True Variance</i>	827
Cureton, Edward E. <i>Communality Estimation in Factor Analysis of Small Matrices</i>	371
Cureton, Edward E. <i>A Measure of the Average Inter-Correlation</i>	627
Damarin, Fred. <i>A Special Review of Buros' Personality Tests and Reviews</i>	215
Doppelt, Jerome E. <i>Differences between the Miller Analogies Test Scores of People Tested Twice</i>	735
Dunnette, Marvin D. (with Richard S. Elster). <i>The Robustness of Tilton's Measure of Overlap</i>	685
Ebel, Robert L. <i>How to Write True-False Test Items</i>	417
Elster, Richard S. (with Marvin D. Dunnette). <i>The Robustness of Tilton's Measure of Overlap</i>	685
Elster, R. S. (with J. R. Capra). <i>A Note on Generating Multivariate Data with Desired Means, Variances, and Covariances</i>	749
Farley, Frank H. (with Herbert H. Severson). <i>The Stability of Individual Differences in Strength and Sensitivity of the Nervous System</i>	453
Feldman, David H. (with Winston Markwalder). <i>Systematic Scoring of Ranked Distractors for the Assessment of Piagetian Reasoning Levels</i>	347
Follman, John (with William Miller and Eldon Burg). <i>Statistical Analysis of Three Critical Thinking Tests</i>	519
Forsyth, Robert A. <i>An Empirical Note on Correlation Coefficients Corrected for Restriction in Range</i>	115
Frary, Jewel M. (with Louis L. McQuitty). <i>Reliable and Valid Hierarchical Classification</i>	321
Frieze, Irene (with Sol M. Roshal and Janet T. Wood). <i>A Multitrait-Multimethod Validation of Measures of Student Attitudes Toward School, Toward Learning, and Toward Technology in Sixth Grade Children</i>	999
Goo, Valerie (with William B. Michael, Robert A. Jones, Hudhail Al Amir, Calvin M. Pullias, and Michel Jackson). <i>Correlates of a Pass-Fail Decision for Admission to Candidacy in a Doctoral Program in Education</i>	965
Goolsby, Thomas M., Jr. (with Donald A. Williamson). <i>Use of the ROTC Qualifying Examination for Selections of Students to Enroll in Advanced Courses in ROTC as Juniors</i>	513
Goolsby, Thomas M., Jr. <i>Appropriateness of Subtests in Achievement Tests Selection</i>	969

Gordon, Leonard V. <i>Are There Two Extremeness Response Sets?</i>	867
Gorth, William P. (with Dwight W. Allen and Aram Grayson). <i>Computer Programs for Test Objectives and Item Banking</i>	245
Grayson, Aram (with William P. Gorth and Dwight W. Allen). <i>Computer Programs for Test Objective and Item Banking</i>	245
Gregory, Thomas B. <i>A Computer Program for the Compilation of Data from Classroom Observation Systems Having Mutually Exclusive Categories</i>	763
Grimaldi, Joseph (with Eugene Loveless, James Hennessy, and John Prior). <i>Factor Analysis of 1970-71 Version of the Comparative Guidance and Placement Battery</i>	959
Gross, Alan L. (with Norman K. Rubin). <i>A Ten Factor Unequal "N" Analysis of Variance Program</i>	753
Guertin, Wilson H. <i>Typing Ships with Transpose Factor Analysis</i>	397
Haemmerlie, Frances M. (with Marvin E. Shaw and J. Thomas Bowman). <i>The Validity of Measures of Eye-Contact</i>	919
Halperin, Silas (with Robert W. Lissitz). <i>A Computer Program for Estimating the Power of Tests of Assumptions of Markov Chains</i>	287
Hamilton, David L. <i>A Comparative Study of Five Methods of Assessing Self-Esteem, Dominance, and Dogmatism</i>	441
Haney, Russell (with William B. Michael, Young B. Lee, and Joan J. Michael). <i>The Criterion-Related Validities of Cognitive and Noncognitive Predictors in a Training Program for Nursing Candidates</i>	983
Harper, Frank B. W. <i>Specific Anxiety Theory and the Mandler-Sarason Test Anxiety Questionnaire</i>	1011
Harris, Chester W. (with Margaret L. Harris). <i>A Factor Analytic Interpretation Strategy</i>	589
Harris, Margaret L. (with Chester W. Harris). <i>A Factor Analytic Interpretation Strategy</i>	589
Hennessy, James (with Joseph Grimaldi, Eugene Loveless, and John Prior). <i>Factor Analysis of 1970-71 Version of the Comparative Guidance and Placement Battery</i>	959
Hewitt, Benjamin A. (with Paul S. Burnham). <i>Advanced Placement Scores: Their Predictive Validity</i>	939
Holmes, David S. <i>The Relationship between Expected Grades and Students' Evaluations of Their Instructions</i>	951
Hooke, Ora (with William B. Michael, Young B. Lee, Joan J. Michael, and Wayne S. Zimmerman). <i>A Partial Redefinition of the Factorial Structure of the Study Attitudes and</i>	

<i>Methods Survey (SAMS) Test</i>	545
Horn, John L. <i>Integration of Concepts of Reliability and Standard Error of Measurement</i>	57
Howell, John R. (with Sharon L. Crews). <i>Eigenvalues and Vectors of Large Matrices on the IBM-1130</i>	263
Howell, Margaret A. <i>Combining the Ipsative and Normative Approaches in Selection Validation</i>	931
Ivens, Stephen H. <i>Nonparametric Item Evaluation Index</i>	843
Jackson, Michel (with William B. Michael, Robert A. Jones, Hudhail Al Amir, Calvin M. Pullias, and Valerie Goo). <i>Correlates of a Pass-Fail Decision for Admission to Candidacy in a Doctoral Program in Education</i>	965
Jacoby, Jacob (with Michael S. Matell). <i>Is There an Optimal Number of Alternatives for Likert-Scale Items? Study I: Reliability and Validity</i>	657
Jones, Robert A. (with William B. Michael, Hudhail Al Amir, Calvin M. Pullias, Michel Jackson, and Valerie Goo). <i>Correlates of a Pass-Fail Decision for Admission to Candidacy in a Doctoral Program in Education</i>	965
Jones, W. Paul (with F. L. Newman). <i>Bayesian Techniques for Test Selection</i>	851
Jung, Steven M. (with Dewey Lipe and Thomas J. Quirk). <i>An Alteration of Program UTEST to Determine the Direction of Group Differences for the Mann-Whitney U Test</i> ..	269
Kalimo, Esko. <i>Notes on Approximate Procrustes Rotation to Primary Pattern</i>	363
Kane, Bernard H. (with Joanne M. Lenke and Harold F. Bligh). <i>Cross-Validation of the Orleans-Hanna Algebra Prognosis Test and the Orleans-Hanna Geometry Prognosis Test</i>	521
Kane, Robert B. <i>Minimizing Order Effects in the Semantic Differential</i>	137
Kane, Robert B. (with William B. Rudolph). <i>A Computer Program for Estimating Relative Sequential Constraint</i> ..	267
Kelly, Robert F. (with Vernon O. Tyler, Jr.). <i>Predicting the Behavior of Institutionalized Delinquents with—and without—Cattell's HSPQ</i>	1019
Klingensmith, John E. (with John W. Menne). <i>Subroutine to Decode IBM 1230 Data</i>	293
Kohr, Richard L. <i>An Item Analysis and Scoring Program for Summated Rating Scales</i>	769
Komorita, S. S. (with Gloria Cowan). <i>The Effects of Forewarning and Pretesting on Attitude Change</i>	431
Kropp, R. P. (with H. W. Stoker). <i>An Empirical Validity Study of the Assumptions Underlying the Structure of Cognitive Processes Using Guttman-Lingoes Smallest Space</i>	

<i>Analysis</i>	469
Krug, Samuel E. (with Raymond B. Cattell). <i>A Test of the Trait-View Theory of Distortion in Measurement of Personality by Questionnaire</i>	721
Labinowich, Edward P. (with James K. Brewer). <i>Computer Programs for Rank Analysis of Covariance</i>	295
Lange, Donald E. <i>An EDP System Package for Scoring the Interpersonal Check List</i>	775
Lee, Young B. (with William B. Michael and Robert A. Smith). <i>The Relationship of Average Scores on Intelligence and Reading Tests to Percentages of Minority Group Students in Elementary Schools and High Schools in a Large Metropolitan Area</i>	539
Lee, Young B. (with William B. Michael, Joan J. Michael, Ora Hooke, and Wayne S. Zimmerman). <i>A Partial Redefinition of the Factorial Structure of the Study Attitudes and Methods Survey (SAMS) Test</i>	545
Lee, Young B. (with William B. Michael, Russell Haney, and Joan J. Michael). <i>The Criterion-Related Validities of Cognitive and Noncognitive Predictors in a Training Program for Nursing Candidates</i>	983
Lenke, Joanne M. (with Harold F. Bligh and Bernard H. Kane). <i>Cross-Validation of the Orleans-Hanna Algebra Prognosis Test and the Orleans-Hanna Geometry Prognosis Test</i>	521
Lessing, Elsie E. (with Susan W. Zagorin). <i>Dimensions of Psychopathology in Middle Childhood as Evaluated by Three Symptom Checklists</i>	175
Lewis, Robert A. <i>A Streamlined Version of the ALDOUS Simulation of Personality</i>	283
Lightsey, Ralph (with Brad S. Chissom). <i>A Comparison of the D-48 Test and the Otis Quick Score for High School Dropouts</i>	525
Lindem, Alfred C. (with John D. Williams). <i>Setwise Regression Analysis—A Stepwise Procedure for Sets of Variables</i>	747
Linn, Robert L. (with Charles E. Werts and Ledyard R. Tucker). <i>The Interpretation of Regression Coefficients in a School Effects Model</i>	85
Linn, Robert L. (with Charles E. Werts). <i>Analyzing School Effects: ANCOVA with a Fallible Covariate</i>	95
Linn, Robert L. (with Charles E. Werts). <i>Considerations When Making Inferences within the Analysis of Covariance Model</i>	407
Linn, Robert L. (with Charles E. Werts). <i>Problems with Inferring Treatment Effects from Repeated Measures</i>	269
Lipe, Dewey (with Steven M. Jung and Thomas J. Quirk). <i>An Alteration of Program UTEST to Determine the Direc-</i>	

<i>tion of Group Differences for the Mann-Whitney U Test . .</i>	269
Lissitz, Robert W. (with Silas Halperin). <i>A Computer Program for Estimating the Powers of Tests of Assumptions of Markov Chains</i>	287
Lord, Frederic M. <i>Robbins-Monro Procedures for Tailored Testing</i>	3
Lord, Frederic M. <i>A Theoretical Study of the Measurement Effectiveness of Flexilevel Tests</i>	805
Loveless, Eugene (with Joseph Grimaldi, James Hennessy, and John Prior). <i>Factor Analysis of 1970-71 Version of the Comparative Guidance and Placement Battery</i>	959
Lu, K. H. <i>A Measure of Agreement among Subjective Judgments</i>	75
Lu, K. H. <i>Statistical Control of "Impurity" in the Estimation of Test Reliability</i>	641
Mann, Lester (with Barton B. Proger, John R. McGowan, Robert J. Bayuk, Jr., Ruth L. Trevorow, and Edward Massa). <i>The Relative Predictive and Construct Validities of the Otis-Lennon Mental Ability Test, the Lorge-Thorndike Intelligence Test, and the Metropolitan Readiness Test in Grades Two and Four: A Series of Multivariate Analyses</i>	529
Markwalder, Winston (with David H. Feldman). <i>Systematic Scoring of Ranked Distractors for the Assessment of Piagetian Reasoning Levels</i>	347
Margolis, Bruce L. (with William H. Clark). <i>A Revised Procedure for the Analysis of Biographical Information</i>	461
Massa, Edward (with Barton B. Proger, John R. McGowan, Robert J. Bayuk, Jr., Lester Mann, and Ruth L. Trevorow). <i>The Relative Predictive and Construct Validities of the Otis-Lennon Mental Ability Test, the Lorge-Thorndike Intelligence Test, and the Metropolitan Readiness Test in Grades Two and Four: A Series of Multivariate Analyses</i>	529
Matell, Michael S. (with Jacob Jacoby). <i>Is There an Optimal Number of Alternatives for Likert Scale Items? Study I: Reliability and Validity</i>	657
McGowan, John, Jr. (with Barton B. Proger, Robert J. Bayuk, Jr., Lester Mann, Ruth L. Trevorow, and Edward Massa). <i>The Relative Predictive and Construct Validities of the Otis-Lennon Mental Ability Test, the Lorge-Thorndike Intelligence Test, and the Metropolitan Readiness Test in Grades Two and Four: A Series of Multivariate Analyses</i>	529
McLaughlin, Gerald W. <i>A Note on the Validity of Two Measures of High School Rank</i>	989
McQuitty, Louis L. <i>Relaxed Rank Order Typal Analysis</i>	33
McQuitty, Louis L. (with Jewel M. Frary). <i>Reliable and Valid Hierarchical Classification</i>	321

McQuitty, Louis L. <i>A Comparative Study of Some Selected Methods of Pattern Analysis</i>	607
McQuitty, Louis L. <i>A Short Cut Toward a Submatrix Containing Only "Disturbed" Individuals</i>	815
Mehrotra, C. M. N. <i>Behavioral Cognition as Related to Interpersonal Perception and Some Personality Traits of College Students</i>	145
Menne, John W. (with John E. Klingensmith). <i>Subroutine to Decode IBM 1230 Data</i>	293
Michael, Joan J. (with William B. Michael, Young B. Lee, Ora Hooke, and Wayne S. Zimmerman). <i>A Partial Redefinition of the Factorial Structure of the Study Attitudes and Methods Survey (SAMS) Test</i>	545
Michael, Joan J. (with William B. Michael, Russell Haney, and Young B. Lee). <i>The Criterion-Related Validities of Cognitive and Noncognitive Predictors in a Training Program for Nursing Candidates</i>	983
Michael, Joan J. (with Maria S. A. Seda). <i>The Concurrent Validity of the Sprigle School Screening Readiness Test for a Sample of Preschool and Kindergarten Children</i>	995
Michael, William B. (with Robert A. Smith and Young B. Lee). <i>The Relationship of Average Scores on Intelligence and Reading Tests to Percentages of Minority Group Students in Elementary Schools and High Schools in a Large Metropolitan Area</i>	539
Michael, William B. (with Young B. Lee, Joan J. Michael, Ora Hooke, and Wayne S. Zimmerman). <i>A Partial Redefinition of the Factorial Structure of the Study Attitudes and Methods Survey (SAMS) Test</i>	545
Michael, William B. (with Robert A. Jones, Hudhail Al Amir, Calvin M. Pullias, Michel Jackson, and Valerie Goo). <i>Correlates of a Pass-Fail Decision for Admission to Candidacy in a Doctoral Program in Education</i>	965
Michael, William B. (with Russell Haney, Young B. Lee, and Joan J. Michael). <i>The Criterion-Related Validities of Cognitive and Noncognitive Predictors in a Training Program for Nursing Candidates</i>	983
Miller, William (with John Follman and Eldon Burg). <i>Statistical Analysis of Three Critical Thinking Tests</i>	519
Newman, F. L. (with W. Paul Jones). <i>Bayesian Techniques for Test Selection</i>	851
Nickel, Ted. <i>The Reduced Size Rod and Frame Test as a Measure of Psychological Differentiation</i>	555
Olsen, LeRoy C. (with William H. Venema). <i>A Projective Occupational Attitudes Test</i>	907
Page, Monte M. <i>Postexperimental Assessment of Awareness in</i>	

<i>Attitude Conditioning</i>	891
Parker, Randall M. <i>A Program of Scheffe's Method</i>	761
Parker, George V. C. <i>Prediction of Individual Stability</i>	875
Parry-Hill, Joseph W., Jr. (with Bert W. Westbrook and Roger W. Woodbury). <i>The Development of a Measure of Vocational Maturity</i>	541
Payne, David A. (with Robert A. Wells and Robert R. Clarke). <i>Another Contribution to Estimating Success in Graduate School: A Search for Sex Differences and Comparison between Three Degree Types</i>	497
Prior, John (with Joseph Grimaldi, Eugene Loveless, and James Hennessy). <i>Factor Analysis of 1970-71 Version of the Comparative Guidance and Placement Battery</i>	959
Proger, Barton B. (with John R. McGowan, Robert J. Bayuk, Jr., Lester Mann, Ruth L. Trevorow, and Edward Massa). <i>The Relative Predictive and Construct Validities of the Otis-Lennon Mental Ability Test, the Lorge-Thorndike Intelligence Test, and the Metropolitan Readiness Test in Grades Two and Four: A Series of Multivariate Analyses</i>	529
Pullias, Calvin M. (with William B. Michael, Robert A. Jones, Hudhail Al Amir, Michel Jackson, and Valerie Goo). <i>Correlates of a Pass-Fail Decision for Admission to Candidacy in a Doctoral Program in Education</i>	965
Quirk, Thomas J. (with Steven M. Jung and Thomas J. Quirk). <i>An Alteration of Program UTEST to Determine the Direction of Group Differences for the Mann-Whitney U Test</i>	269
Ramsay, J. O. <i>True Score Theory: A Paradox</i>	715
Reeb, M. <i>A One-Step Nomograph for the Kolmogorov-Smirnov Test</i>	887
Reiner, John R. <i>Differential Effects of Initial Course Placement as a Function of ACT Mathematics Scores and High School Rank-in-Class in Predicting General Performance in Chemistry</i>	977
Renzulli, Joseph S. (with Robert A. Shaw). <i>The Reliability and Validity of the Contemporary Mathematics Test</i>	973
Richards, James M., Jr. <i>Prediction of Choice of and Success in Agriculture as a College Major</i>	505
Roberge, James J. <i>A Computer Program for Nonparametric Post Hoc Comparisons for Trend</i>	275
Roberge, James J. <i>A Computer Program for Trend Analysis in a Two- or Three-Factor Experiment with Repeated Measures on One of the Factors</i>	279
Roberge, James J. <i>A Computer Program for Nonparametric Post Hoc Multiple Comparisons</i>	755
Rohlf, Richard J. <i>A Higher-Order Alpha Factor Analysis of</i>	

- Interest, Personality, and Ability Variables, Including an Evaluation of the Effect of Scale Interdependency* 381
- Roshal, Sol M. (with Irene Frieze and Janet T. Wood). *A Multitrait-Multimethod Validation of Measures of Student Attitudes Toward School, Toward Learning, and Toward Technology in Sixth Grade Children* 999
- Rubin, Norman K. (with Alan L. Gross). *A Ten Factor Unequal "N" Analysis of Variance Program* 753
- Rudolph, William B. (with Robert B. Kane). *A Computer Program for Estimating Relative Sequential Constraint* 267
- Schmidt, Frank L. *The Relative Efficiency of Regression and Simple Unit Predictor Weights in Applied Differential Psychology* 699
- Schnitzen, Joseph P. (with John A. Cox). *Concurrent Validity of a Literature Test in Relation to Selection of Persons for Graduate Study in English* 485
- Schuldt, David L. (with Robert F. Stahmann). *Interest Profiles of Clergymen as Indicated by the Vocational Preference Inventory* 1025
- Seda, Maria S. A. (with Joan J. Michael). *The Concurrent Validity of the Sprigle School Screening Readiness Test for a Sample of Preschool and Kindergarten Children* 995
- Severson, Herbert H. (with Frank H. Farley). *The Stability of Individual Differences in Strength and Sensitivity of the Nervous System* 453
- Sharon, Amiel T. *Measurement of College Achievement by the College-Level Examination Program* 477
- Shaw, Marvin E. (with J. Thomas Bowman and France M. Haemmerlie). *The Validity of Measures of Eye-Contact ..* 919
- Shaw, Robert A. (with Joseph S. Renzulli). *The Reliability and Validity of the Contemporary Mathematics Test* 973
- Shine, Lester C. II (with Samuel M. Bower). *A One-Way Analysis of Variance for Single-Subject Designs* 105
- Siegelman, Marvin. *SAT and High School Average Predictions of Four Year College Achievement* 947
- Smith, I. Leon. *Validity of Taxonomic Tests* 475
- Smith, Robert A. (with William B. Michael and Young B. Lee). *The Relationship of Average Scores on Intelligence and Reading Tests to Percentages of Minority Group Students in Elementary Schools and High Schools in a Large Metropolitan Area* 539
- Stahmann, Robert F. (with David L. Schuldt). *Interest Profiles of Clergymen as Indicated by the Vocational Preference Inventory* 1025
- Stoker, H. W. (with R. P. Kropp). *An Empirical Validity Study of the Assumptions Underlying the Structure of Cog-*

<i>nitive Processes Using Guttman-Lingoes Smallest Space Analysis</i>	460
Tatham, Clifford B. (with Elaine J. Tatham). <i>A Note on the Predictive Validity of the Cooperative Algebra III</i>	517
Tatham, Elaine J. (with Clifford B. Tatham). <i>A Note on the Predictive Validity of the Cooperative Algebra III</i>	517
Terranova, Carmelo. <i>Factor Similarity</i>	261
Thomas, Jerry R. (with Brad S. Chissom). <i>Multivariate Validity of the Otis Lennon Mental Ability Tests Primary I Level</i>	991
Trevorrow, Ruth L. (with Barton B. Proger, John R. McGowan, Robert J. Bayuk, Jr., Lester Mann, and Edward Massa). <i>The Relative Predictive and Construct Validities of the Otis-Lennon Mental Ability Test, the Lorge-Thorn-dike Intelligence Test, and the Metropolitan Readiness Test in Grades Two and Four: A Series of Multivariate Analyses</i>	529
Tucker, Ledyard R. (with Robert L. Linn and Charles E. Werts). <i>The Interpretation of Regression Coefficients in a School Effects Model</i>	85
Tyler, Vernon O., Jr. (with Robert F. Kelly). <i>Predicting the Behavior of Institutionalized Delinquents with—and without—Cattell's HSPQ</i>	1019
Venema, William H. (with LeRoy C. Olsen). <i>A Projective Occupational Attitudes Test</i>	907
Waters, Carrie Wherry (with A. G. Bayroff). <i>A Comparison of Computer-Simulated Conventional and Branching Tests</i>	125
Waters, Carrie Wherry (with Lawrence K. Waters). <i>Validity and Likability Ratings for Three Scoring Instructions for a Multiple-Choice Vocabulary Test</i>	935
Wells, Robert A. (with David A. Payne and Robert R. Clarke). <i>Another Contribution to Estimating Success in Graduate School: A Search for Sex Differences and Comparison between Three Degree Types</i>	497
Welsh, George S. <i>Vocational Interests and Intelligence in Gifted Adolescents</i>	155
Werts, Charles E. (with Robert L. Linn and Ledyard R. Tucker). <i>The Interpretation of Regression Coefficients in a School Effects Model</i>	85
Werts, Charles E. (with Robert L. Linn). <i>Analyzing School Effects: ANCOVA with a Fallible Covariate</i>	95
Werts, Charles E. (with Robert L. Linn). <i>Considerations When Making Inferences within the Analysis of Covariance Model</i>	407
Werts, Charles E. (with Robert L. Linn). <i>Problems with Inferring Treatment Effects from Repeated Measures</i>	857
Westbrook, Bert W. (with Joseph W. Parry-Hill, Jr. and Roger W. Woodbury). <i>The Development of a Measure of</i>	

<i>Vocational Maturity</i>	541
Wiggins, Nancy. <i>Individual Differences in Diagnostic Judgments of Psychosis and Neurosis from the MMPI</i>	199
Williams, John D. (with Alfred C. Lindem). <i>Setwise Regression Analysis—A Stepwise Procedure for Sets of Variables</i>	747
Williamson, Donald A. (with Thomas M. Goolsby, Jr.). <i>Use of the ROTC Qualifying Examination for Selection of Students to Enroll in Advanced Courses in ROTC as Juniors</i> ..	513
Wood, Janet T. (with Sol M. Roshal and Janet T. Wood). <i>A Multitrait-Multimethod Validation of Measures of Student Attitudes Toward School, Toward Learning, and Toward Technology in Sixth Grade Children</i>	999
Woodbury, Roger W. (with Bert W. Westbrook and Joseph W. Parry-Hill, Jr.). <i>The Development of a Measure of Vocational Maturity</i>	541
Zagorin, Susan W. (with Elsie E. Lessing). <i>Dimensions of Psychopathology in Middle Childhood as Evaluated by Three Symptom Checklists</i>	175
Zimmerman, Wayne S. (with William B. Michael, Young B. Lee, Joan J. Michael, and Ora Hooke). <i>A Partial Redefinition of the Factorial Structure of the Study Attitudes and Methods Survey (SAMS) Test</i>	545



U. S. POSTAL SERVICE
STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION
(Act of August 12, 1970: Section 3685, Title 39, United States Code)

DATE OF FILING

September 26, 1971.

TITLE OF PUBLICATION

Educational and Psychological Measurement

FREQUENCY OF ISSUE

Quarterly

LOCATION OF KNOWN OFFICE OF PUBLICATION (Street, city, county, state, zip code)

61 Byrdhill Road, Richmond, Virginia 23228

LOCATION OF THE HEADQUARTERS OR GENERAL BUSINESS OFFICES OF THE PUBLISHER (Not printers)

31 Cheek Road, Durham, N. C. 27704

NAMES AND ADDRESSES OF PUBLISHER, EDITOR, AND MANAGING EDITOR

PUBLISHER (Name and address)

Frederic Kuder, Box 6907 College Station, Durham, N. C. 27708

EDITOR (Name and address)

W. Scott Gehman, Box 6907 College Station, Durham, N. C. 27708

MANAGING EDITOR (Name and address)

Geraldine R. Thomas, 3121 Cheek Road, Durham, N. C. 27704

OWNER (If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by partnership or other unincorporated firm, its name and address, as well as that of each individual must be given.)

NAME

Frederic Kuder (Owner)

ADDRESS

Box 6907 College Station, Durham, N. C. 27708

KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES (If there are none, so state)

None

	AVERAGE NO. COPIES EACH ISSUE DURING PRECEDING 12 MONTHS	ACTUAL NUMBER OF COPIES OF SINGLE ISSUE PUBLISHED NEAREST TO FILING DATE
EXTENT AND NATURE OF CIRCULATION		
TOTAL NO. COPIES PRINTED (Net Press Run)	3435	3509
PAID CIRCULATION		
1. SALES THROUGH DEALERS AND CARRIERS		
STREET VENDORS AND COUNTER SALES		
2. MAILING SUBSCRIPTIONS	2697	2778
TOTAL PAID CIRCULATION	2697	2778
FREE DISTRIBUTION (including samples) BY MAIL		
CARRIER OR OTHER MEANS	30	30
TOTAL DISTRIBUTION (Sum of C and D)	2727	2808
OFFICE USE, LEFT-OVER, UNACCOUNTED, SPOILED AFTER PRINTING	708	701
TOTAL (Sum of E & F—should equal net press run shown in A)	3435	3509

I certify that the statements made by me above are correct and complete.

(Signature of editor, publisher, business manager, or owner)

Geraldine R. Thomas, Managing Editor

